Universidade Federal de Campina Grande
Centro de Engenharia Elétrica e Informática
Coordenação de Pós-Graduação em Ciência da Computação

Analisando padrões de mobilidade a partir de redes sociais e de dados sociodemográficos abertos

Caio Libânio Melo Jerônimo

Campina Grande, Paraíba, Brasil

Universidade Federal de Campina Grande
Centro de Engenharia Elétrica e Informática
Coordenação de Pós-Graduação em Ciência da Computação

Analisando padrões de mobilidade a partir de redes sociais e de dados sociodemográficos abertos

Caio Libânio Melo Jerônimo

Dissertação submetida à Coordenação do Curso de Pós-Graduação em Ciência da Computação da Universidade Federal de Campina Grande – Campus I como parte dos requisitos necessários para obtenção do grau de Mestre em Ciência da Computação.

Área de Concentração: Ciência da Computação
Linha de Pesquisa: Banco de Dados e Descoberta da Informação

Cláudio Elízio Calazans Campelo, Ph.D.
(Orientador)

Cláudio de Souza Baptista, Ph.D.
(Orientador)

Campina Grande, Paraíba, Brasil

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA CENTRAL DA UFCG

J56a Jerônimo, Caio Libânio Melo.

Analisando padrões de mobilidade a partir de redes sociais e de dados sociodemográficos abertos / Caio Libânio Melo Jerônimo. – Campina Grande, 2017.

103 f. : il. color.

Dissertação (Mestrado em Ciência da Computação) – Universidade Federal de Campina Grande, Centro de Engenharia Elétrica e Informática, 2017.

"Orientação: Prof. Dr. Cláudio Elízio Calazans Campelo, Prof. Dr. Cláudio de Souza Baptista".

Referências.

1. Análise de Correlação. 2. Dados Abertos. 3. Dados Espaciais e Temporais. 4. Padrões de Mobilidade. 5. Redes Sociais. I. Campelo, Cláudio Elízio Calazans. II. Baptista, Cláudio de Souza. III. Título.

CDU 004.051(043)

"ANALISANDO PADRÕES DE MOBILIDADE A PARTIR DE REDES SOCIAIS E DE DADOS SOCIODEMOGRÁFICOS ABERTOS"

CAIO LIBÂNIO MELO JERÔNIMO

DISSERTAÇÃO APROVADA EM 07/07/2017

CLAUDIO ELIZIO CALAZANS CAMPELO, PhD., UFCG
Orientador(a)

CLÁUDIO DE SOUZA BAPTISTA, PhD., UFCG
Orientador(a)

LEANDRO BALBY MARINHO, Dr., UFCG
Examinador(a)

JOÃO PORTO DE ALBUQUERQUE, Dr., UK
Examinador(a)

CAMPINA GRANDE - PB

Este trabalho é dedicado à memória de meu pai,
Neucimar Jerônimo Leite.

Agradecimentos

Agradeço aos meus pais, Magna e Neucimar, por todo o esforço ao longo de suas vidas para me criar, e sobretudo para me educar, mostrando que a educação é a única esperança de um futuro digno para nossa sociedade. Sem educação, restam apenas paus e pedras.

Agradeço a minha noiva, Elayne, pela paciência que teve ao longo de todo o mestrado, suportando minhas inúmeras ausências, posto que essas foram necessárias para a conclusão deste trabalho.

Aos meus orientadores, o professor Cláudio Elízio Calazans Campelo e o professor Cláudio de Souza Baptista, pela confiança, orientação neste trabalho, paciência e por todo o conhecimento que compartilharam, o qual levarei sempre comigo.

Aos meus amigos de longa data, pelos momentos de descontração e amizade que proporcionaram, e também pelo apoio dado.

Agradeço aos colegas do Laboratório de Sistemas de Informação pelas experiências trocadas e pelo apoio, como também pela infraestrutura fornecida pelo laboratório para o desenvolvimento apropriado deste trabalho.

Sou grato aos professores e demais funcionários da Universidade Federal de Campina Grande, do Centro de Engenharia Elétrica e Informática, do Departamento de Sistemas e Computação e da Coordenação de Pós-Graduação em Ciência da Computação que apoiaram de forma direta ou indireta a realização deste trabalho.

Agradeço à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pelo apoio financeiro prestado.

"...tenho a impressão de ter sido uma
criança brincando à beira-mar, divertindo-me em
descobrir uma pedrinha mais lisa ou uma concha mais
bonita que as outras, enquanto o imenso oceano da verdade
continua misterioso diante de meus olhos."
(Isaac Newton)

Resumo

A demanda constante por melhorias na qualidade de vida dos habitantes das grandes cidades, somado à crescente urbanização desses centros, torna imprescindível a utilização de meios tecnológicos para um melhor entendimento da dinâmica dos centros urbanos e como seus habitantes interagem nesses ambientes. Nesse sentido, o aumento na utilização de dispositivos eletrônicos equipados com sistemas GPS e o constante anseio da humanidade por comunicação e, mais atualmente, por conexão à internet, vem criando novas oportunidades de estudo e também grandes desafios, especialmente no que tange a grande quantidade de dados gerados pelas redes sociais. Diversas pesquisas vêm utilizando esses dados para realizar estudos que buscam compreender traços do comportamento humano, especialmente no que diz respeito à mobilidade urbana e trajetórias. Porém, grande parte das pesquisas que utilizam dados georreferenciados se restringem às dimensões espaciais e temporais, desconsiderando outros aspectos que podem influenciar na mobilidade humana. Este trabalho propõe um método computacional capaz de extrair padrões de mobilidade oriundos de mensagens georreferenciadas de redes sociais e correlacioná-los com indicadores sociais, econômicos e demográficos fornecidos por órgãos governamentais, buscando assim, analisar quais possíveis fatores poderiam exercer alguma influência sobre a mobilidade dos moradores de uma grande cidade. Para validar o método proposto, foram utilizadas mensagens postadas no Twitter e um conjunto de indicadores sociais, ambos oriundos da cidade de Londres. Os resultados mostraram a existência de correlações entre padrões de mobilidade e indicadores sociais, especialmente os relacionados com condições de emprego e renda, como também com características étnico-religiosas dos indivíduos em estudo.

Palavras-chave: análise de correlação. dados abertos. dados espaciais e temporais. padrões de mobilidade. redes sociais.

Abstract

The constant need for improvements in life quality of inhabitants of big cities, together with the increasing urbanization of these centers, demands the use of technological means for a better understanding of the dynamics of urban centers and how their inhabitants interact in these environments. In this sense, the adoption of electronic devices equipped with GPS systems, the human need for communication and, more recently, for Internet connection, have brought new research opportunities and great challenges, especially due to the huge amount of data generated by social networks. Several studies have used this data to carry out research that seek to understand traces of human behavior, especially with respect to urban mobility and trajectories. However, much of the research that uses georeferenced data are restricted to spatial and temporal dimensions, disregarding other aspects that may influence human mobility. This work proposes a model capable of extracting mobility patterns from georeferenced messages of social networks and correlating them with social, economic and demographic indicators provided by government agencies, seeking to analyze which factors may impact in urban mobility. To evaluate the model, we used messages posted on Twitter and a set of social indicators, both related to the city of London. The results revealed the existence of correlations between mobility patterns and social indicators, especially those related to employment and income conditions, as well as ethnic and religious characteristics of the individuals under study.

Keywords: correlation analysis. mobility patterns. open data. social networks. spatio-temporal data.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

AC	*Activity center*
API	*Application Programming Interface*
DBSCAN	*Density-based spatial clustering of applications with noise*
GPS	*Global Positioning System*
LSOA	*Lower Super Output Area*
POI	*Point of interest*
RFID	*Radio-Frequency Identification*
SOM	*Self-organizing map*

Lista de símbolos

ρ	Letra grega minúscula rho
τ	Letra grega minúscula tau

Sumário

Parte I

Introdução

1 Introdução

Com o crescimento das grandes metrópoles e com a constante necessidade de melhorias nos serviços prestados aos seus habitantes, vem crescendo a demanda por serviços que permitam, em especial aos órgãos públicos, entender a dinâmica das cidades e como estas se relacionam com seus habitantes. Neste contexto, tendências de pesquisas científicas vêm surgindo, especialmente no âmbito de cidades inteligentes, permitindo assim o uso da tecnologia para melhorar a infraestrutura urbana, proporcionando uma melhor qualidade de vida a seus habitantes.

Batty et al. (2012) destacam que a um dos principais componentes relacionados ao crescente interesse no estudo da dinâmica urbana está relacionado à grande produção de dados (*big data*), em especial os associados à mobilidade urbana. Estes dados permitem a identificação de padrões de mobilidade, os quais expressam características do comportamento humano, ajudando em estudos relacionados a políticas de transportes públicos, segurança pública, engenharia de tráfego e demais aspectos associados ao planejamento de cidades (NOULAS et al., 2012; WILSON; BELL, 2004).

Dados associados à mobilidade urbana podem ser coletados por meio de tecnologias wireless, como sistemas de posicionamento global e também redes de telefonia móvel. Porém, é importante destacar que o uso massivo de redes sociais, bem como a popularização no uso de celulares modernos (que, em sua maioria, estão equipados com GPS) vêm permitindo estudos mais aprofundados no tocante à mobilidade urbana, favorecendo também o uso destes dados em sistemas de recomendação (HAO et al., 2010; ZHENG et al., 2010), bem como em estudos abordando trajetórias (BAGROW; LIN, 2012; HSIEH; LI; LIN, 2012).

Atualmente, diversos sistemas online, redes sociais e aplicativos móveis permitem o compartilhamento de informações relacionadas à localização atual do usuário, assim como a postagens de checkins associados a determinados pontos de interesse (POI), trazendo assim mais informações, como, por exemplo, dados relacionados aos preços destes POI, popularidade do local, e até mesmo informações semânticas extraídas dos textos das próprias postagens realizadas pelos usuários.

Ainda no contexto de grandes cidades, é sabido que estas frequentemente apresentam diversas discrepâncias, especialmente nas esferas econômicas, sociais e demográficas onde, em uma mesma cidade, pode-se encontrar muitas variações nesses aspectos, como, por exemplo, regiões mais ricas e mais pobres, regiões com um maior índice de pessoas imigrantes e também regiões com uma maior concentração de pessoas. Analisar como estes fatores podem influenciar os padrões de mobilidade de uma população constitui um grande desafio a ser considerado, tanto na própria obtenção destes indicadores sociais, quanto na

extração de padrões de mobilidade da população.

Diante do exposto, esta dissertação propõe um método computacional capaz de extrair padrões de mobilidade oriundos de mensagens de redes sociais, e de correlacionar estes padrões com dados sociais, econômicos e demográficos[1]. Para tal, o método proposto trata mensagens coletadas da rede social Twitter[2], permitindo a extração de propriedades estatísticas que descrevem padrões de mobilidade, correlacionando assim, estes padrões com dados sociais fornecidos ao método.

O método proposto inclui componentes responsáveis por: filtrar mensagens relevantes à análise; detectar locais de residência dos usuários que postaram as mensagens; detectar regiões frequentemente visitadas pelos usuários (chamados de *Activity Centers* - AC) e também os POI visitados; extrair padrões de mobilidade das mensagens georreferenciadas; calcular as correlações entre padrões de mobilidade e os dados sociais fornecidos ao método; e gerar da matriz de correlação. Para a extração de padrões de mobilidade, são considerados indicadores frequentemente utilizados na literatura, bem como são propostas novas métricas de mobilidade para a extração destes padrões. Detalhes sobre estas métricas são apresentadas no Capítulo 4 deste trabalho.

A cidade de Londres (Reino Unido) foi utilizada para estudo de caso, sendo coletado um total de 19.456.798 mensagens postadas no Twitter para esta região. Para os dados sociais, utilizou-se a plataforma *London Datastore*[3] como fonte de dados. Esta plataforma concentra diversos indicadores relacionados às regiões da cidade de Londres, permitindo a exibição destes dados em forma de gráficos e mapas, facilitando assim o estudo analítico destas informações tanto para fins científicos, como para a própria população desta região.

Diversos desafios foram encontrados no decorrer desta pesquisa, em especial os relacionados à natureza dos dados colhidos do Twitter, os quais, em sua grande maioria, estão incompletos ou fragmentados. Por exemplo, um usuário pode postar mensagens apenas de sua casa e de seu trabalho, ocasionando uma visão limitada dos seus padrões de mobilidade, fazendo-se necessária a implementação de técnicas para a mitigação de problemas relacionados aos dados.

Através dos experimentos executados, foram identificadas diversas correlações entre padrões de mobilidade e dados sociais, em especial no tocante a indicadores sociais relacionados a condições de emprego, populações estrangeiras e condições de renda da população.

[1] Por simplicidade, neste trabalho, os dados sociais, econômicos e demográficos são referidos apenas como "dados sociais".

[2] Twitter: <https://twitter.com/>

[3] London Datastore: <http://data.london.gov.uk/>

1.1 Definição do problema

Diversas pesquisas têm analisado padrões de mobilidade em regiões urbanas, em sua maioria utilizando dados coletados de redes de celulares (GONZALEZ; HIDALGO; BARABASI, 2008; JIANG et al., 2013; PALCHYKOV et al., 2014), redes Wifi (CHAINTREAU et al., 2007; ZHANG et al., 2012) e sinais de GPS (RHEE et al., 2011; ZHAO et al., 2014). Muitos desses trabalhos apresentam restrições relacionadas aos dados utilizados como, por exemplo, a baixa precisão de coordenadas coletadas de redes de telefonia móvel, ou mesmo o uso de um pequeno número de voluntários para a coleta de dados de mobilidade.

Além das restrições supracitadas, as pesquisas na área não abordam de forma consistente possíveis relacionamentos entre padrões de mobilidade e outros dados que não estejam vinculados às esferas espaciais e temporais. Essa característica será discutida no Capítulo 3, que apresenta os trabalhos relacionados, onde diversos autores destacam a ausência de diferentes tipos de dados nos estudos de mobilidade, em especial, dados relacionados a indicadores sociais, econômicos e demográficos de uma população. Dado esse contexto, a pesquisa apresentada nesta dissertação visa desenvolver um método computacional capaz de integrar variáveis de mobilidade e indicadores sociais, permitindo uma análise mais ampla sobre suas possíveis correlações. Para a realização desta análise, a pesquisa implementa técnicas para a detecção de padrões de mobilidade presentes em mensagens georreferenciadas do Twitter, bem como utiliza um vasto conjunto de indicadores sociais oriundos de censos demográficos.

1.2 Objetivos

1.2.1 Objetivos Gerais

O principal objetivo desta pesquisa é criar um método computacional que permita identificar padrões de mobilidade de pessoas a partir de mensagens georreferenciadas de redes sociais, e identificar correlações estatísticas entre estas informações de mobilidade e indicadores socioeconômicos da região oriundos de plataformas de dados abertos.

1.2.2 Objetivos Específicos

1. **Desenvolver método para a filtragem e seleção de mensagens georreferenciadas do Twitter:** desenvolver métodos que permitam filtrar e selecionar mensagens que sejam apropriadas ao estudo. Por exemplo, mensagens do Twitter que não possuam coordenadas geográficas, ou que estas coordenadas estejam fora dos limites geográficos da região em estudo deverão ser removidas. Também deverão ser desconsideradas mensagens relacionadas a usuários que postam poucas mensagens, assim como aqueles que postam apenas de uma única localização.

2. **Implementar técnica para a detecção de residências dos usuários a partir das suas mensagens postadas publicamente na rede social:** implementar uma solução para a detecção das residências dos usuários que postam mensagens na rede social. Esta informação é útil pois a região da residência de um indivíduo, em geral, está associada a diversos indicadores sociais, sendo esta informação necessária para a análise de correlação entre padrões de mobilidade de um indivíduo e a região em que ele reside.

3. **Implementar técnicas para a detecção de regiões e POI frequentemente visitados pelos usuários:** implementar métodos para a detecção de regiões frequentemente visitadas por um indivíduo. Esta detecção deverá se basear nas mensagens postadas com frequência em uma mesma localização, permitindo identificar também os POI que este indivíduo mais frequenta. Este estudo se faz relevante para analisar como os padrões de mobilidade identificados pelo método se correlacionam com os lugares e regiões que o indivíduo mais frequenta.

4. **Implementar métricas de mobilidade existentes:** Implementar métricas de mobilidade que sejam utilizadas em trabalhos na literatura para descrever padrões de mobilidade de pessoas.

5. **Desenvolver novas métricas de mobilidade:** Elaborar e implementar métricas de mobilidade que ainda não foram apresentadas na literatura, favorecendo assim uma análise mais detalhada entre as possíveis interações deste tipo de dado com os indicadores sociais utilizados na pesquisa.

6. **Implementar solução em software capaz de identificar padrões de mobilidade e correlacioná-los com indicadores sociais:** desenvolver um método que, a partir das mensagens georreferenciadas coletadas e filtradas, bem como das regiões de residência, AC e POI detectados, possa extrair padrões de mobilidade e correlacioná-los com dados socioeconômicos.

7. **Validar o método proposto:** conduzir uma avaliação experimental para validação do método proposto, utilizando a cidade de Londres (Reino Unido) e a rede social Twitter como estudo de caso.

1.3 Contribuições

As principais contribuições apresentadas por este trabalho são:

1. Desenvolvimento de um método capaz de extrair padrões de mobilidade a partir de mensagens georreferenciadas. Esta funcionalidade pode ser estendida para diversas

outras aplicações que trabalhem com estes padrões, sendo assim um facilitador para pesquisas em diversos outros campos de estudo.

2. Desenvolvimento de método computacional capaz de integrar e correlacionar padrões de mobilidade extraídos de redes sociais e indicadores sociais, econômicos e demográficos de uma região.

3. Uso conjunto de diversas técnicas presentes na literatura, como detecção de residências, detecção de AC e POI com o objetivo de extrair padrões de mobilidade de usuários de redes sociais.

4. Identificação de parâmetros que auxiliam a extração de padrões de mobilidade de redes sociais.

5. Implementação de técnicas de filtragem de mensagens de redes sociais, objetivando eliminar mensagens pouco representativas.

1.4 Trabalhos publicados

O seguinte artigo foi publicado contendo uma descrição de parte do método desenvolvido e alguns resultados preliminares:

> JERÔNIMO, C. L. M; CAMPELO, C. E. C.; BAPTISTA, C. S. Analyzing mobility patterns from social networks and social, economic and demographic open data. In: *Proceedings of the XVII Brazilian Symposium on Geoinformatics (GeoInfo 2016)*. Campos do Jordão, SP, Brazil, 2016. pp 32-43, ISSN 2179-4820.

O artigo foi um dos seis trabalhos selecionados para publicação no *Journal of Information and Data Management* (JIDM), tendo uma versão estendida submetida para a revista. Outro artigo está sendo escrito para submissão a uma revista científica internacional, apresentando mais detalhes sobre o método desenvolvido e os resultados obtidos.

Ainda no escopo das atividades desenvolvidas ao longo do mestrado, sendo parte da disciplina de Fundamentos de Pesquisa em Ciência da Computação (FPCC), foi publicado o seguinte trabalho:

> JERÔNIMO, C. L. M.; CAMPELO, C. E. C.; BAPTISTA, C. S. Mining influential terms for toponym recognition and resolution. *Revista Brasileira de Cartografia*, v. 68, n. 6, 2016.

O trabalho supracitado teve como objetivo o aprimoramento de um *Geoparser*, visando a melhoria na detecção de toponímias presentes em documentos de texto.

1.5 Organização estrutural

Os demais capítulos desta dissertação estão organizados da seguinte forma: o Capítulo 2 apresenta a fundamentação teórica para este trabalho, mostrando os principais pontos para o entendimento de assuntos relacionados à detecção de padrões de mobilidade, bem como as técnicas utilizadas para a agregação de mensagens (*clustering*) e para o cálculo das correlações estatísticas. O Capítulo 3 apresenta um levantamento de trabalhos mais representativos no tocante a padrões de mobilidade, com foco nos trabalhos que manipulam dados de redes sociais. O Capítulo 4 descreve o método proposto, as técnicas utilizadas e detalhes de suas implementações. No Capítulo 5, descrevem-se todos os experimentos realizados para a validação do método descrito, bem como os resultados obtidos e limitações encontradas no trabalho. Finalmente, o Capítulo 6 apresenta as considerações finais e apontamentos para trabalhos futuros.

Parte II

Fundamentação teórica

2 Fundamentação teórica

Para a realização do estudo de padrões de mobilidade, faz-se necessário considerar, desde técnicas de mineração de dados, a conceitos que envolvem o desenvolvimento de tecnologias que já fazem parte do dia a dia das pessoas, como é o caso das redes sociais.

Buscando favorecer um entendimento geral sobre temáticas que permeiam esta pesquisa, este capítulo visa abordar temas e conceitos que servem de base para a construção dos métodos aqui empregados.

O presente capítulo está organizado com a seguinte estrutura: a Seção 2.1 descreve os conceitos de descoberta da informação e mineração de dados, suas características e aplicações. A Seção 2.2 aborda a temática de redes sociais e como pesquisas de análise de dados são empregadas nestas redes para permitir a extração de conhecimento. A Seção 2.3 descreve os principais métodos de análise de correlação entre variáveis. Por fim, a Seção 2.4 apresenta as considerações finais do capítulo.

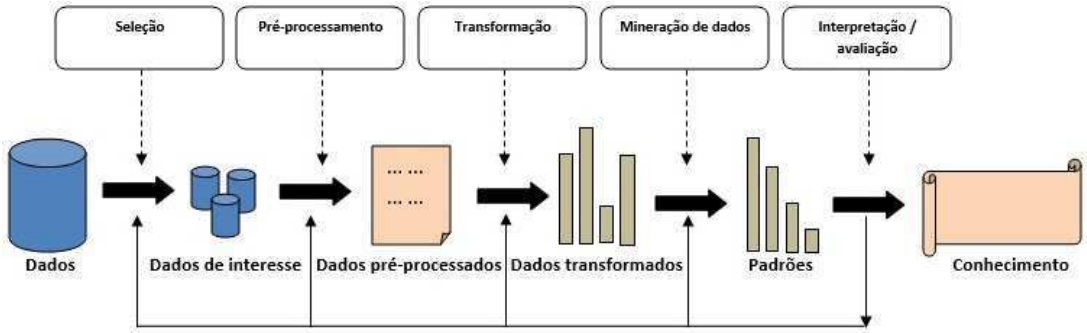
2.1 Descoberta da informação e mineração de dados

Com o desenvolvimento de tecnologias que permitem uma maior participação das pessoas no tocante à produção de informação, em especial às tecnologias associadas ao conceito de Web 2.0, tornou-se imperativo o desenvolvimento de ferramentas computacionais capazes de processar o enorme volume de informações gerados todos os dias. Neste contexto, técnicas associadas à descoberta da informação e mineração de dados vêm ganhando espaço no cotidiano das pessoas, se tornando conceitos imprescindíveis no mundo atual.

A descoberta da informação, em inglês *Knowledge Discovery* (KDD) tem como principal objetivo o desenvolvimento de métodos e técnicas que permitam extrair informações úteis de grandes volumes de dados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996), enquanto mineração de dados consiste no processo de extrair padrões dos dados, sendo então, de acordo com Wei, Piramuthu e Shaw (2003), a mineração de dados uma das etapas da descoberta de informação. A Figura 1 mostra as etapas presentes no processo de descoberta da informação.

A primeira etapa (Seleção) consiste em selecionar dados ou amostras de dados que serão analisadas com o objetivo de encontrar alguma informação útil, onde estes dados geralmente são originados em diferentes bases de dados. A segunda etapa (Pré-processamento) lida com os problemas associados à integração das informações (e.g. dados originados em diferentes bancos de dados) e também realiza uma limpeza nestes dados,

Figura 1 – Etapas do processo de descoberta da informação



Fonte: Adaptado de Wei, Piramuthu e Shaw (2003, p. 158)

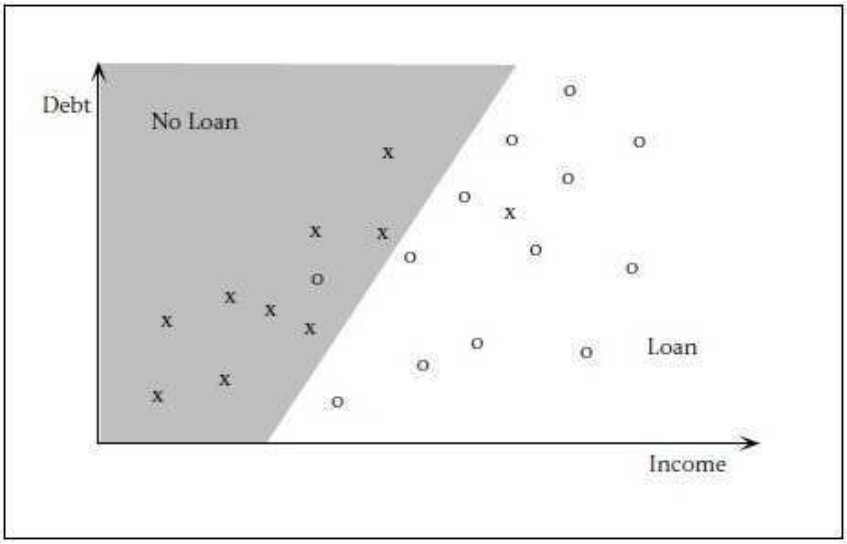
removendo, por exemplo, possíveis *outliers*. A terceira etapa (Transformação) transforma as informações da etapa anterior em dados que possam ser interpretados pelas técnicas de mineração de dados utilizada. A quarta etapa (Mineração de dados) extrai padrões presentes nos dados. Por último, os padrões encontrados são analisados e interpretados quanto à sua possível utilidade, podendo, inclusive, retroceder o processo para qualquer uma das etapas anteriores, caso seja necessário (WEI; PIRAMUTHU; SHAW, 2003).

Segundo Fayyad, Piatetsky-Shapiro e Smyth (1996), os dois principais objetivos da mineração de dados são, em geral, a predição e a descrição. A predição consiste em utilizar variáveis conhecidas para prever valores de variáveis de interesse. Já a descrição, consiste em encontrar padrões que possam ser interpretados de alguma forma.

Como um dos métodos de mineração mais utilizados, temos a classificação, que permite atribuir uma determinada classe a uma informação. Este método pode ser usado, por exemplo, para classificação do nível de risco para empréstimos financeiros. A Figura 2 ilustra este exemplo, onde é demonstrado duas classes de usuários. Os usuários identificados como "x" são aqueles que atrasaram parcelas de algum empréstimo, e os usuários classificados como "o" pagam suas parcelas em dia. Baseado em diversas informações, o banco pode estabelecer uma partição simples entre os usuários, separando assim, os clientes que estariam habilitados a receber um empréstimo (*Loan*) e os usuários que não poderiam recebe-lo (*No Loan*). É importante observar que a classificação entre usuários que podem ou não receber o empréstimo não é perfeita, porém, o modelo de classificação do banco consegue distinguir boa parte dos usuários com sucesso (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

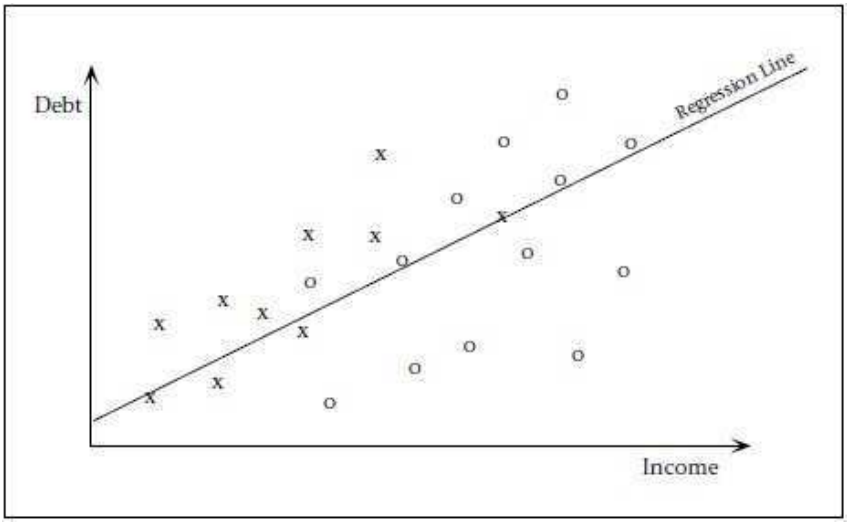
O segundo método é a regressão, que permite realizar previsões para valores de uma determinada variável. A Figura 3 mostra uma regressão linear simples, onde o total de débitos está representado como uma função linear da renda dos clientes de um banco. Com este modelo, o banco pode, por exemplo, prever a quantidade de débitos que um cliente

Figura 2 – Classificação linear de usuários como hábeis ou não para a tomada de empréstimos



Fonte: Fayyad, Piatetsky-Shapiro e Smyth (1996, p. 44)

Figura 3 – Regressão linear simples entre total de débitos e renda dos clientes de um banco



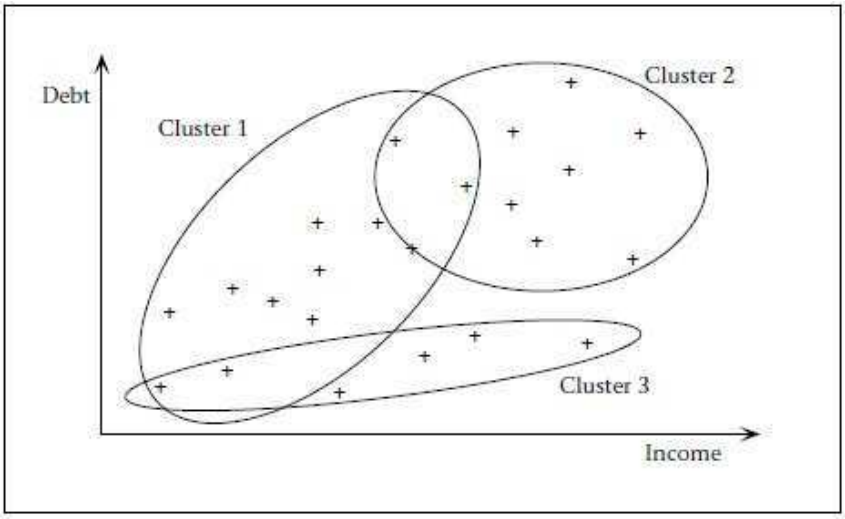
Fonte: Fayyad, Piatetsky-Shapiro e Smyth (1996, p. 44)

terá baseado unicamente nos dados de sua renda (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

O terceiro método é a agregação, que consiste em encontrar *clusters* entre os dados, onde os elementos de cada *cluster* possuem um grau de semelhança entre si, respeitando uma métrica de distanciamento (WEI; PIRAMUTHU; SHAW, 2003). Este método pode ser utilizado, por exemplo, para a detecção de eventos que estejam ocorrendo uma cidade,

de acordo com o grau de aglomeração das pessoas em uma área específica. A Figura 4 mostra dados agrupados em diferentes *clusters*. No exemplo, os usuários identificados como "x" e "o" foram substituídos pelo símbolo "+" indicando não conhecimento destes usuários, sabendo-se apenas o necessário para agrupá-los em um dos *clusters* no exemplo.

Figura 4 – Dados agrupados em diferentes *clusters*



Fonte: Fayyad, Piatetsky-Shapiro e Smyth (1996, p. 45)

O quarto método é a sumarização, que consiste em sumarizar um conjunto de valores em um único valor. Um exemplo para este método seria o cálculo da média de um conjunto de valores. Técnicas mais sofisticadas consistem, por exemplo, em derivação de regras e descoberta de relacionamentos funcionais entre variáveis (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). O último método, segundo Fayyad, Piatetsky-Shapiro e Smyth (1996), é a modelagem de dependências, que consiste em encontrar um modelo que descreve dependências significativas entre variáveis. Análises de correlações estatísticas entre variáveis são exemplos deste método de mineração de dados.

Como demonstrado, existem inúmeras aplicações que podem utilizar algumas das técnicas de mineração de dados para descobrir determinados padrões presentes nas informações analisadas. O uso destas técnicas se faz especialmente necessária quando se considera as informações geradas por grandes empresas, onde, cada nova descoberta pode representar uma vantagem estratégica, permitindo, por exemplo, o aprimoramento de serviços prestados, a contensão de despesas e descoberta de novas aplicações para produtos.

2.2 Redes sociais e análise de dados

Segundo Kaplan e Haenlein (2010), mídias sociais se constituem de grupos de aplicações baseadas na internet que, possuindo ideologias fundadas na Web 2.0, permite a

participação dos usuários na criação e compartilhamento de conteúdo.

A massificação no uso de redes sociais, especialmente após a popularização de dispositivos móveis por parte da população, propiciou uma acentuação nas características de criação e compartilhamento destas redes. Atualmente, a maioria dos dispositivos móveis vendidos no mercado já saem de fábrica capazes de se conectarem a redes de acesso à internet, bem como podem utilizar sinais de GPS para calcular suas localizações em terra. Isto faz com que praticamente qualquer usuário possa se conectar à internet e compartilhar experiências vividas, postando, por exemplo, fotos georreferenciadas na rede.

Atualmente, existem diversas redes sociais com os mais diversos objetivos, possibilitando a agregação de usuários com interesses em comum, ou mesmo permitindo o compartilhamento de fotos e textos na rede. Barbier e Liu (2011) apresenta na Tabela 1 uma categorização entre as principais aplicações disponíveis para estas redes. Já a Figura 5 demonstra o crescimento no número de usuários de redes sociais ao longo dos últimos anos. Os dados foram extraídos da plataforma *Pew Research Center*[1]. Esse rápido crescimento no número de usuários demonstra a massificação no uso destas redes, o que favorece o desenvolvimento de pesquisas que utilizam estes dados como base para seus estudos.

Tabela 1 – Categorização de redes sociais

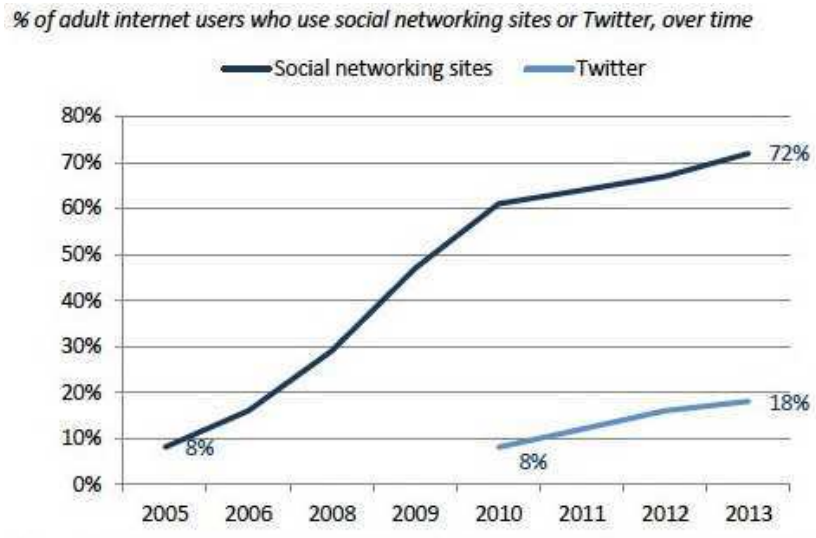
Categoria	**Exemplos**
Blogs	Blogger, LiveJournal, WordPress
Microblogs	Twitter, GoogleBuzz
Opinion mining	Epinions, Yelp
PhotoandvideoSharing	Flickr, YouTube
Social bookmarking	Delicious, StumbleUpon
Social networking sites	Facebook, LinkedIn, MySpace, Orkut
Social news	Digg, Slashdot
Wikis	Scholarpedia, Wikihow, Wikipedia, Event maps

Fonte: Barbier e Liu (2011)

No contexto de análise de dados de redes sociais, técnicas de mineração de dados podem ser empregadas para uma melhor compreensão dos padrões que estão presentes neste tipo de informação. Usos deste tipo de análise podem ser observadas em aplicações como detecção de tópicos em mensagens (XIE et al., 2016; MIAO et al., 2016; ZHANG et al., 2016), detecção de eventos (SAKAKI; OKAZAKI; MATSUO, 2010; ADEDOYIN-OLOWE et al., 2016), análise de sentimentos (BARBOSA; FENG, 2010; SAIF et al., 2016), análise e previsão de trajetórias (GABRIELLI et al., 2014; NETO; BAPTISTA;

[1] Pew Research Center's Internet & American Life Project: <http://www.pewinternet.org/2013/08/05/72-of-online-adults-are-social-networking-site-users/>

Figura 5 – Uso de redes sociais e do Twitter ao longo do tempo



Fonte: Pew Research Center's Internet & American Life Project

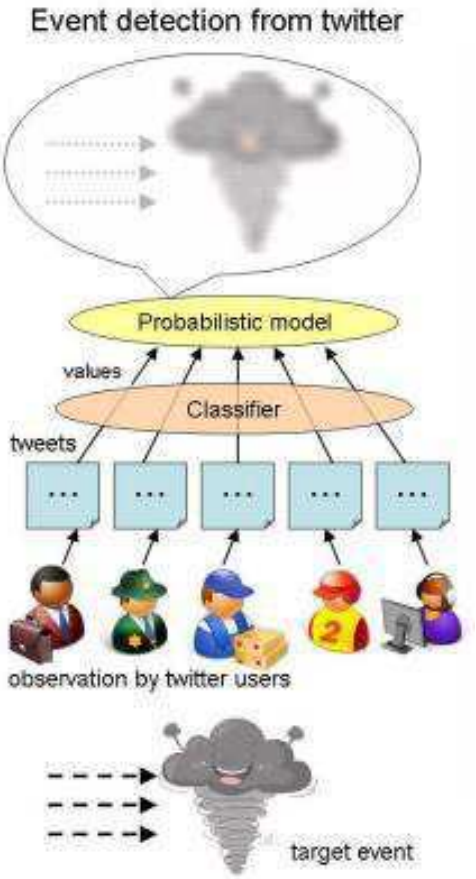
CAMPELO, 2016) e demais aplicações que envolvam a detecção de padrões existentes nos dados gerados nestas redes.

Em um clássico estudo sobre detecção de eventos por meio de dados do Twitter, Sakaki, Okazaki e Matsuo (2010) investigam as interações de usuários da rede social em relação a eventos ocorrendo em tempo real, como terremotos e trajetória de furacões, e propõem um algoritmo capaz de monitorar e detectar tais eventos. Os autores desenvolvem um classificador baseado em palavras chave encontrada nas mensagens, no número de palavras e seus respectivos contextos, bem como desenvolvem um modelo probabilístico capaz de traçar a trajetória do evento baseado nas mensagens colhidas. A Figura 6 mostra o processo de detecção de eventos por meio de mensagens postadas no Twitter, utilizando os próprios usuários desta rede como sensores. Neste processo, as mensagens postadas pelos usuários são submetidas a um classificador, que ao analisar o conteúdo das mensagens, atribui valores para cada uma delas, classificando-as como mensagens que estejam relacionadas a algum evento ou não. Com base nos resultados apresentados pelo classificador, o modelo probabilístico desenvolvido pode detectar, de fato, o possível evento em curso.

Focando ainda em eventos relacionados a desastres naturais, Albuquerque et al. (2015) apresentam uma abordagem para identificação de mensagens do Twitter que sejam relevantes em um contexto de enchentes ou inundações. O trabalho considera, além das próprias mensagens georreferenciadas da rede social, dados geológicos coletados de órgãos oficiais, tendo como estudo de caso, a enchente do Rio Elba em 2013 na Alemanha. Os resultados demonstram que mensagens postadas em uma distância de até 10km de áreas

inundadas possuem maiores chances de estar relacionadas a este tipo de evento.

Figura 6 – Processo de detecção de eventos por meio de mensagens do twitter



Fonte: Adaptado de Sakaki, Okazaki e Matsuo (2010, p. 854)

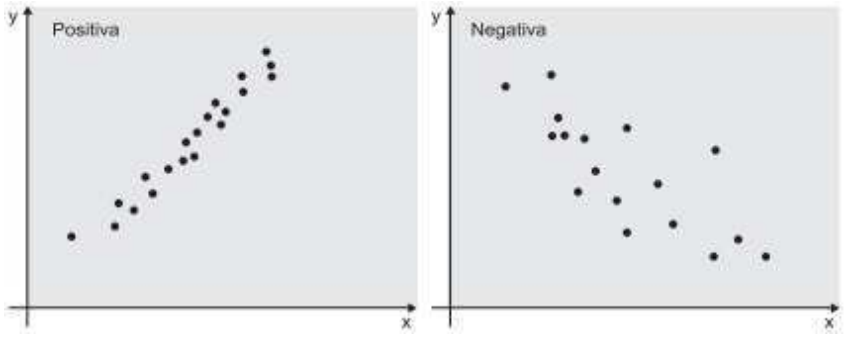
2.3 Análise de correlação estatística

Coeficientes de correlação são valores que representam a associação entre duas variáveis. Estes coeficientes são de extrema importância, especialmente em estudos científicos que buscam entender, tanto a direção, quanto a força destas associações, não significando, necessariamente, relações de causa e efeito, mas sim uma tendência de variação que as variáveis apresentam.

Estudos envolvendo correlação de variáveis estão presentes em diversas áreas de pesquisa, como em trabalhos envolvendo políticas públicas de saúde, pesquisas nas áreas de engenharia, ciências médicas, bem como em grande parte de pesquisas que buscam entender este tipo de associação.

Os coeficientes de correlação mais utilizados na literatura são os de Pearson, Spearman e Kendall, sendo os dois últimos utilizados para dados não-normais (CHOK,

Figura 7 – Direção das correlações estatísticas entre duas variáveis



Fonte: Naghettini e Pinto (2007, p. 357)

2010). Seus valores variam de -1 a 1, onde valores negativos indicam correlações negativas (inversas) entre duas variáveis, e valores positivos indicam correlações positivas. Quanto mais distante do zero, maiores são as correlações entre as variáveis. Na Figura 7 são apresentados exemplos de correlações positivas e negativas.

Dentre os três coeficientes de correlação mais utilizados na literatura, o de Pearson é o mais comum, tendo como principais características: (1) utilizado para medir associação linear entre variáveis; (2) as variáveis precisam ter distribuição aproximadamente normal; (3) necessita considerar a homoscedasticidade dos dados, que consiste no grau de dispersão das variâncias em relação à reta de regressão; (4) necessita especial controle ou eliminação de *outliers*. A Equação 2.1 (CHOK, 2010) demonstra a formalização do coeficiente de correlação de Pearson, onde x_i e y_i são os valores para cada par de variáveis e n é o número de pares.

$$\rho = \frac{\sum_{i=1}^{n}((x_i - \overline{x})(y_i - \overline{y}))}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{2.1}$$

Onde:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n};\ \overline{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

O coeficiente de correlação de Spearman é uma medida de correlação não-paramétrica, isto é, não assume que os dados das amostras sigam uma distribuição específica, costumando ser utilizado quando as prerrogativas inerentes ao teste de Pearson são violadas, sendo o coeficiente de Spearman baseado em *ranking*. O teste de Spearman pode ser utilizado em casos de não-normalidade dos dados e também para medir relacionamentos não lineares entre as variáveis. A Equação 2.2 (ZAR, 1972) demonstra a formalização do coeficiente de Spearman, onde d^2 é a diferença entre as ordenações.

$$\rho_s = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n^3 - n} \tag{2.2}$$

O terceiro coeficiente de correlação é o de Kendall que, assim como o de Spearman, é um método não-paramétrico baseado em *ranking*, sendo este, em sua variação *tau-b*, mais resistente a repetições nos dados presentes nas amostras. A Equação 2.3 demonstra a formalização para o coeficiente *tau* de Kendall, onde n_c é o número de pares concordantes e n_d é o número de pares discordantes.

$$\tau = \frac{n_c - n_d}{\frac{1}{2} n(n-1)} \tag{2.3}$$

2.4 Considerações finais

Neste capítulo, foram apresentados os principais conceitos que estão relacionados ao desenvolvimento deste trabalho. Aqui foram discutidas questões relacionadas ao processo de descoberta da informação, onde o processo de mineração de dados constitui-se de uma das etapas que levam a este descobrimento, permitindo a extração de padrões encontrados nestes dados.

Também foram discutidas questões relacionadas ao uso de redes sociais para a extração de informação, as principais categorias deste tipo de rede, e as possibilidades de análise de seus dados. É importante ressaltar que, devido ao crescimento no uso destas redes, desafios relacionados ao tratamento de grandes volumes de informações vem sendo cada vez mais discutidos na literatura, trazendo à tona conceitos como, por exemplo, o de *big data* e seus impactos nos processos de descoberta da informação.

O próximo capítulo irá apresentar uma revisão literária acerca de trabalhos relacionados à análise de padrões de mobilidade a partir de redes sociais.

Parte III

Trabalhos relacionados

3 Trabalhos relacionados

Nos últimos anos, com a escalada na produção de informações pela humanidade em praticamente todas as áreas de conhecimento, surgiu uma crescente necessidade por técnicas que possam encontrar informações úteis nos dados gerados. Atualmente, estima-se que cerca de 2,5 quintilhões de bytes são produzidos diariamente pela humanidade[1], evidenciando o grande desafio a ser enfrentado no que toca as áreas da descoberta da informação, mineração de dados e *big data*.

Boa parte do crescimento associado à produção de informação está, sem dúvidas, relacionado à popularização das redes sociais, bem como ao uso de dispositivos móveis que permitem, por exemplo, gerar informações georreferenciadas a partir de praticamente qualquer lugar em que o usuário possa estar.

Neste contexto, diversas pesquisas vêm sendo realizadas nos últimos anos buscando utilizar informações georreferenciadas para descobrir determinados comportamentos humanos, especialmente os associados a deslocamentos em centros urbanos, podendo assim, gerar uma vasta gama de conhecimentos sobre como estes indivíduos se comportam neste espaço.

Contudo, inúmeros desafios ainda se fazem presentes em estudos que se propõem a analisar dados georreferenciados associados à mobilidade urbana. Muitos estudos utilizam dados obtidos de sinais emitidos por antenas de celulares, onde estas informações podem possuir problemas relacionados à precisão geográfica. Também é comum observar na literatura uma vasta gama de estudos que analisam padrões de mobilidade restringindo-se unicamente a aspectos espaciais e temporais, não verificando como estes podem ser estudados frente a outras questões pertinentes ao comportamento humano. Outra questão relevante acerca de estudos relacionados a padrões de mobilidade, é a carência de trabalhos que se propõem a desenvolver métodos automatizados para a análise destes padrões, permitindo uma abordagem mais prática sobre o problema.

Desta forma, foi desenvolvida uma revisão da literatura objetivando identificar as principais contribuições de trabalhos que estudam padrões de mobilidade, bem como determinar o estado da arte acerca deste tema.

As próximas seções deste capítulo estão organizadas como segue: a Seção 3.1 apresenta uma revisão literária dos trabalhos que se propõem a extrair e analisar padrões de mobilidade de dados obtidos de redes sociais, considerando as dimensões espaciais e temporais apenas. A Seção 3.2 busca estender o conteúdo apresentado na seção anterior,

[1] Tome nota: 2,5 quintilhões de bytes são criados todos os dias. Disponível em: <http://cio.com.br/noticias/2015/10/27/tome-nota-2-5-quintilhoes-de-bytes-sao-criados-todos-os-dias/>

apresentando trabalhos que levem em consideração aspectos sociais em suas análises de mobilidade. A Seção 3.3 apresenta uma sumarização dos trabalhos relacionados de acordo com as métricas de mobilidade mais utilizadas. A Seção 3.4 sumariza de forma geral, considerando as principais características dos trabalhos apresentados, facilitando a compreensão do estado da arte. A Seção 3.5 encerra o capítulo apresentando as considerações finais.

3.1 Extração e análise de padrões de mobilidade em redes sociais

Existem na literatura diversas pesquisas relacionadas a padrões de mobilidade, onde muitas destas pesquisas utilizam dados provenientes de redes de celulares, dispositivos RFID e redes Wifi, porém, muitos destes trabalhos possuem restrições relacionadas à precisão das coordenadas geográficas. Zandbergen (2009) ao estudar a precisão associada aos sistemas de A-GPS, sinais WiFi e de antenas de celulares, observou erros de aproximadamente 9 metros para A-GPS, 74 metros para redes WiFi e de até 962 metros para sinais de celulares.

Apenas recentemente, trabalhos abordando o tema de mobilidade foram desenvolvidos utilizando dados de redes sociais para traçar e delimitar padrões de mobilidade. Estes trabalhos possuem o benefício da larga adoção de sistemas de GPS em dispositivos móveis, favorecendo assim a precisão das coordenadas geográficas obtidas. Portanto, neste capítulo, maior atenção será direcionada a estes trabalhos.

No trabalho de Yuan et al. (2013), é proposto um modelo probabilístico denominado W^4 (Who + Where + When + What) para extrair, a partir de mensagens do Twitter, características da mobilidade de usuários do ponto de vista espacial, temporal e também das atividades realizadas por estes. Neste trabalho, os autores modelam as interações destes quatro fatores, objetivando identificar comportamentos dos indivíduos por meio da descoberta de tópicos e interesses destes usuários em função do tempo e dos espaços geográficos visitados. Os experimentos realizados pelos autores demonstram o bom desempenho do modelo em capturar padrões de comportamento. No trabalho, também é demonstrado o interesse dos autores em considerar informações de cunho social em estudos futuros.

Wakamiya, Lee e Sumiya (2011) estudam as características de mobilidade de usuários do Twitter localizados no Japão. Os autores extraem as características de mobilidade baseado em três indicadores: (1) total de mensagens postadas dentro de uma área urbana; (2) o número total de usuários distintos presentes em uma área urbana e (3) o número de usuários que se deslocam de uma área para outra dentro do ambiente urbano. Com estas métricas, os autores categorizam as áreas urbanas do Japão como sendo predominantemente cidades residenciais (*bedroom town*); cidades com atividades comerciais intensas (*office town*); cidades com uma maior atividade noturna, contendo

muitos bares e restaurantes (*nightlife towns*); e regiões mistas (*multifunctional towns*).

Jurdak et al. (2015) desenvolvem um estudo sobre *tweets* georreferenciados originados na Austrália, com o objetivo de demonstrar que estas mensagens podem expressar padrões reais de mobilidade humana, tanto dentro de uma cidade, quanto entre diferentes cidades. Os autores utilizam métricas como a distribuição de deslocamento e o Raio de Giro para permitir a classificação dos usuários de acordo com a distância que estes percorrem baseado nas suas mensagens georreferenciadas. Também foi aferida a taxa de previsibilidade da sequência de mensagens postadas pelos usuários, com o objetivo de entender melhor a relação entre padrões de mobilidade e comportamento na rede social. Como resultados, os autores relatam que as mensagens georreferenciadas podem prover características e resultados tanto quanto outras fontes de informação, como por exemplo de antenas de celulares, demonstrando que as mensagens do Twitter podem ser uma boa base para estudo sobre mobilidade urbana. Os autores ressaltam também a necessidade em se considerar dados demográficos em trabalhos futuros.

Chen, Chiang e Peng (2016) buscam descobrir padrões presentes na movimentação diária de usuários do Gowalla e Brightkite, permitindo traçar a evolução destes padrões por meio de agregações nos dados e por meio de dois algoritmos propostos, chamados de GreedyKL e GreedyMDL, que têm como objetivo extrair padrões de mobilidade que descrevem os movimentos dos usuários em função do tempo. O GreedyKL particiona um conjunto de dados de *check-in* em uma sequência de segmentos em um intervalo de tempo, permitindo se obter informações sobre a evolução nos deslocamentos, enquanto o GreedyMDL combina os segmentos extraídos utilizando princípios do MDL (*Minimum Description Length*). Com o *framework* proposto, é possível retornar como saída os padrões de movimentação mais representativos nos dados analisados.

No trabalho de Hawelka et al. (2014), os autores analisam *tweets* de diversos países para traçar e identificar padrões de mobilidade globais presentes nestas mensagens georreferenciadas. No estudo, são utilizadas métricas como Raio de Giro e diversidade de destinos entre países, comparando assim os padrões de mobilidade existentes entre pessoas de diferentes regiões do globo. Os autores apresentam gráficos demonstrando a densidade de deslocamento de pessoas de diferentes regiões, e também apresentam dados mostrando que o uso de mensagens georreferenciadas do Twitter pode ser utilizado como um *proxy* para o estudo de padrões de mobilidade em escala global, mesmo considerando (o que ficou evidente no trabalho) a existência de viés entre os diferentes países, onde em países mais ricos, as pessoas parecem tender a viajar mais. Como trabalhos futuros, os autores destacam o interesse em estudar padrões de mobilidade em escalas menores (cidades e bairros).

O trabalho de Hasan, Zhan e Ukkusuri (2013) faz uma análise sobre padrões de mobilidade urbana de dados coletados do Twitter. Os autores categorizam os padrões

encontrando as distribuições das atividades ao redor de uma cidade, permitindo determinar os propósitos específicos de uma dada atividade realizada. Para tal, eles fazem uso de *links* referentes ao Foursquare, permitindo identificar e categorizar os *check-ins* como: (1) Em casa; (2) No trabalho; (3) Em refeição; (4) Atividade de entretenimento; (5) Recreação e (6) Fazendo compras. Os autores utilizam as cidades de Nova York, Chicago e Los Angeles no estudo, e as representam como um conjunto de células (200x200 metros) e caracterizaram as movimentações em função da visita de usuários em células específicas, permitindo gerar mapas de distribuição de atividades para cada uma das células. Como conclusões, os autores descrevem que os usuários não escolhem suas rotas de forma randômica, mas tendem a visitar os locais mais populares a outros usuários.

Yin et al. (2015) propõem um modelo probabilístico, denominado *Topic-Region-Model* (TRM) para descobrir, simultaneamente, a semântica, padrões temporais e espaciais de *check-ins* relacionados às atividades dos usuários, permitindo modelar a decisão destes por determinados pontos de interesse (POI). Neste trabalho, são utilizados dados do Foursquare e Twitter, onde são considerados para a recomendação dos POI dados referentes às atividades associadas aos *check-ins* realizados pelos usuários e o tópico ou assunto ao qual um usuário provavelmente está se referindo na sua mensagem. Os experimentos realizados demonstram a eficácia do modelo proposto, especialmente quando são realizadas recomendações em regiões onde o usuário não costuma visitar. Em uma abordagem semelhante, Ferrari et al. (2011) busca extrair tópicos de mensagens do Twitter, permitindo a identificação de *hotspots* presentes na cidade de Nova York. Com o trabalho, é possível identificar locais da cidade que apresentam maiores fluxos de pessoas durante os dias da semana e finais de semana, caracterizando atividades de trabalho e de lazer, respectivamente.

Birkin et al. (2014) buscam classificar os usuários do Twitter de acordo com seus padrões de mobilidade, levando em consideração usuários de regiões urbanas e rurais da cidade de Leeds. Os autores subdividem as mensagens seguindo blocos temporais (início da manhã, final da manhã, tarde e noite) para cada dia da semana, permitindo analisar os usuários de acordo com o tempo que estes permanecem em suas residências, o nível geral de atividade na rede social e o balanço de mensagens ao longo dos dias da semana. Com o estudo, os autores subdividiram os usuários de acordo com grupos de mobilidade distintos, sendo estes grupos: *Family and Friends, Local Hobbyists, Homemaker, Neighbour, Socialite, Student, Executives, Commuter.*

Blanford et al. (2015) investiga os padrões de mobilidade entre as fronteiras políticas de regiões do Kenia, levando em consideração aspectos temporais (dias e meses) e espacial (local e nacional). Os *tweets* foram coletados e filtrados, sendo então mapeados por meio do ArcGIS, permitindo a visualização dos dados coletados ao redor das diferentes regiões políticas do país, levando em conta os *footprints* deixados pelos usuários. Os autores

criaram séries temporais para analisar as mensagens, contendo: (1) movimentação diária dos usuários de acordo com os *tweets* postados, (2) movimentação mensal dos usuários, e (3) movimentação geral (agregando dias e meses) de cada usuário. Os autores utilizam o Raio de Giro dos usuários, tendo como centro de massa o local de maior frequência de postagem das mensagens, bem como calculam o grau de conectividade dos nós (cidades), indicando a frequência em que estes nós são visitados. Como conclusão, os autores relatam a eficiência do uso do Twitter para traçar padrões de mobilidade entre cidades e distritos do Kenia, e realçam a importância deste tipo de estudo em áreas como controle de disseminação de doenças e no estudo da dinâmica e estrutura de comunidades.

Propondo uma abordagem que considere características de mobilidade e círculos de amizade nas redes sociais, Nguyen e Szymanski (2012) utilizaram dados coletados da rede social Gowalla para criar e validar modelos de mobilidade humana levando em consideração estes círculos de amizades e como isto afetaria estes padrões de mobilidade. Os autores fazem a análise com base em três dimensões principais: distância, afinidade e tempo. A distância refere-se à máxima distância entre um *check-in* e outro de um usuário, afinidade está relacionada com a frequência em que um usuário faz *check-in* em uma mesma localização e o tempo é o *timestamp* de um *check-in*, que é utilizado para estimar o tempo entre um *check-in* e outro (quão rápido um usuário se move de um local para outro), bem como o número de postagens em sequência na mesma localização. Os autores verificaram que os círculos de amizade tendem a diminuir com o aumento da distância física entre os indivíduos. Para pesquisas posteriores, os autores destacam a necessidade em se analisar os impactos que fatores econômicos poderiam ter sobre os padrões de mobilidade.

Noulas et al. (2012) utilizam dados do Foursquare para fazer uma análise dos padrões de mobilidade urbana em diferentes cidades do mundo, levando em consideração apenas aspectos espaciais e temporais, com o objetivo de verificar se os padrões de movimentação são similares entre indivíduos de diferentes cidades. Para a análise, considerou-se apenas a probabilidade de deslocamento de um ponto a outro nas cidades por parte dos usuários. Neste estudo, os autores chegam à conclusão que as variações na mobilidade encontrada entre as cidades estudadas (Huston, São Francisco e Singapura) se deram principalmente devido a diferenças em aspectos espaciais entre estas cidades, como diferentes distribuições nas localizações de lugares.

Dredze et al. (2016) descrevem como mensagens do Twitter podem ser utilizadas para a análise de padrões de mobilidade globais, focando seu estudo em deslocamentos realizados ao redor do mundo, e verificando o nível de uso do Twitter em diferentes regiões do planeta. No trabalho, são utilizadas ferramentas externas, como o Geonames[2] para a extração de informações adicionais sobre lugares presentes nas mensagens. Os experimentos executados demonstram um baixo uso da rede social em países mais pobres,

[2] Geonames: <http://geonames.org>

bem como destacam a importância do uso do conteúdo das mensagens para extração de dados semânticos.

Steiger et al. (2016) propõem uma abordagem combinando *self-organizing map* (SOM) e Geo-SOM, que consistem em redes neurais artificiais que produzem mapas bidimensionais a partir de propriedades multidimensionais passadas como entrada. O principal objetivo do estudo é realizar uma análise e comparação de dados oficiais relativos a eventos de tráfego urbano da cidade de Londres (acidentes, congestionamentos e demais eventos) com mensagens georreferenciadas do Twitter que estejam inseridas dentro do mesmo contexto semântico. Como principais resultados, os autores identificaram correlações significativas entre os eventos de trânsito presentes nos dados oficiais e os identificados nas mensagens georreferenciadas da rede social, especialmente para as categorias relacionadas a eventos especiais, incidentes de trânsito e situações de perigo.

3.2 Extração e análise de padrões de mobilidade em redes sociais considerando aspectos sociais

A grande maioria das pesquisas que estudam padrões de mobilidade extraídos de redes sociais focam-se, majoritariamente, nos aspectos espaciais e temporais destes padrões. Porém, outras variáveis podem exercer alguma influência nestes padrões de mobilidade, especialmente fatores sociais inerentes às populações das grandes cidades.

Cranshaw et al. (2012) desenvolvem um algoritmo de agregação que permite subdividir a cidade de Pittsburgh em diferentes *clusters*. Este algoritmo baseia-se nas localizações dos *check-ins* realizados pelos usuários do Foursquare, permitindo aos pesquisadores estabelecer contrapontos às subdivisões políticas da cidade. Com isso, verificou-se que bairros mais pobres obtiveram pouca representatividade nos *clusters*, demonstrando que estes usuários, possivelmente pelas suas condições de renda, possuem pouco acesso à dispositivos móveis com sistemas GPS, se comparado a usuários de outras regiões da cidade.

Cheng et al. (2011) realizam uma pesquisa ampla, com o objetivo de investigar mensagens georreferenciadas do Twitter considerando, além dos aspectos espaciais e temporais, variáveis relacionadas à renda, popularidade na rede social, bem como o conteúdo das mensagens, relacionando a ocorrência de determinadas palavras aos padrões de mobilidade demonstrados. Neste estudo, são utilizadas as métricas de Raio de Giro, distância entre deslocamentos e probabilidade e retorno. Como resultado, especialmente no tocante à renda dos indivíduos, os autores concluem que pessoas que vivem em cidades com uma renda média mais alta, tendem a se locomover por distâncias maiores. Infelizmente, a análise abordada no trabalho é bastante superficial no tocante a este indicador social, não considerando, por exemplo, bairros ou sub-regiões das cidades.

O trabalho de Luo et al. (2016) tem como objetivo principal investigar os padrões de mobilidade de usuários do Twitter, considerando aspectos espaciais e temporais, bem como características demográficas associadas aos usuários da rede. Os autores realizaram o estudo levando em consideração as características relacionadas a etnia dos usuários, a idade e sexo. Para a obtenção destas dados, os autores utilizaram os nomes dos usuários presentes no perfil da rede social, de forma a poder inferir, com uso de informações extraídas de censos, qual a etnia, sexo e idade das pessoas. Além disso, também foram detectados o local das residências dos mesmos. Após o estudo, foi possível observar que, dentre as três variáveis analisadas (etnia, sexo e idade), a que apresentou maior variação nos padrões de mobilidade urbana na cidade de Chicago foi a relacionada à etnia, evidenciando assim possíveis segregações, principalmente relacionado a estrangeiros e imigrantes presentes nesta cidade.

Também considerando questões socioeconômicas em seu estudo, Li, Goodchild e Xu (2013) estuda correlações entre densidade de postagens do Twitter e Flickr e dados sociais e econômicos de municípios do estado da Califórnia. Neste estudo, as densidades de postagens são extraídas e correlacionadas com indicadores sociais extraídos de censos demográficos, onde são consideradas cinquenta e oito variáveis agrupadas nas seguintes categorias: idade, raça, nível de escolaridade, renda e ocupação profissional. Como resultado, os pesquisadores descobriram que pessoas com um bom nível de educação, trabalhando em áreas administrativas e empresariais, científicas e artistas tendem a gerar mais conteúdo georreferenciado, tanto no Twitter como no Flickr.

Steiger et al. (2015) exploram a semântica de mensagens postadas pelo Twitter na região da grande Londres, de modo a inferir as suas possíveis localizações. Nesta pesquisa, os tópicos são extraídos e, de acordo com o tópico de cada mensagem, ela é agrupada como sendo originária de uma residência ou do local de trabalho, permitindo aos pesquisadores, por exemplo, identificar as regiões da cidade onde existe uma maior prevalência de mensagens relacionadas a cada um destes dois grupos. Com estas informações, os autores tentam correlacionar estas regiões com dados de censos que indiquem regiões residenciais ou de trabalho. Os resultados reportaram correlações apenas para as regiões de trabalho, demonstrando a eficácia da extração de tópicos para este caso. Para as regiões de residência, não foram observadas correlações significativas, o que foi atribuído à maior complexidade na detecção de tópicos no contexto de residências.

Gong (2016) analisa como a escolha de uma determinada rede social está relacionada com as atividades executadas pelos usuários, bem como com características demográficas dos mesmos. Para a extração de informações demográficas, tais como sexo, raça e idade, a pesquisa utiliza uma ferramenta de detecção de faces onde, ao processar as imagens dos perfis dos usuários, a ferramenta retorna, de acordo com a análise facial, as informações demográficas associadas à imagem. Nesta pesquisa, são utilizadas as redes sociais Twitter,

Instagram, Foursquare e Weibo com mensagens coletadas das cidades de Rotterdam e Shenzhen. Os autores identificaram que o Twitter e Weibo possuem características de postagens associadas a usuários locais, onde os tópicos possuem um caráter mais geral, com uma maior participação de pessoas jovens e de meia-idade no Twitter e de jovens no Weibo. Em contraste, o Instagram demonstrou ser usado mais por pessoas que escrevem em diferentes idiomas (possivelmente não-locais), apresentou mais tópicos sobre fatos específicos e seu uso prevaleceu em pessoas de meia-idade e jovens.

3.3 Sumarização de métricas de mobilidade

Na literatura, pesquisas que estudam padrões de mobilidade o fazem de diversas formas, utilizando uma grande variedade de atributos e propriedades de mobilidade, não existindo uma única maneira para a realização de tal análise. A Tabela 2 sumariza os trabalhos relacionados de acordo com as métricas de mobilidade utilizadas.

Dentre os trabalhos analisados, foram consideradas as seguintes propriedades, onde: (1) o Raio de Giro representa o desvio padrão entre as mensagens postadas e seu centro de massa (localização média de onde as mensagens são postadas); (2) a distância entre mensagens postadas representa a distância entre mensagens consecutivas postadas pelos usuários; (3) Probabilidade de deslocamentos representa, de forma genérica, a probabilidade de um usuário retornar a um determinado local após um determinado tempo; (4) a autocorrelação espacial permite analisar o grau de semelhança entre objetos que estejam geograficamente próximos; (5) uso de técnicas de agregação para a análise de mobilidade; (6) a taxa de deslocamento entre diferentes países permite analisar o grau de deslocamento de indivíduos entre diferentes países; (7) a densidade de postagem de mensagens representa a intensidade de *posts* por usuários, ou mesmo por regiões geográficas; (8) a quantidade de usuários que postam mensagens em uma região geográfica específica.

A sumarização apresentada na Tabela 2 deixa evidente as diversas métricas de mobilidade adotadas nos trabalhos. O estudo proposto nesta dissertação, que pode ser visto na última linha desta tabela, tem como objetivo utilizar, além do Raio de Giro, distância geográfica entre mensagens e técnicas de agregação, propor novas métricas de mobilidade que permitam refinar esta análise e, em conjunto com as métricas citadas, proporcionar um melhor entendimento na dinâmica da mobilidade urbana. As métricas de mobilidade adotadas nesta pesquisa são descritas em detalhes no Capítulo 4 deste trabalho.

Tabela 2 – Sumarização das métricas de mobilidade utilizadas nos trabalhos analisados

Trabalho	Raio de Giro	Distância entre mensagens	Probabilidade de deslocamentos	Autocorrelação espacial	Técnicas de agregação	Tráfego urbano	Taxa de deslocamento entre países	Densidade de postagens	Quantidade de usuários
Gong (2016)	x								
Luo et al (2016)	x				x				
Steiger et al. (2016)				x	x	x			
Steiger et al. (2015)				x	x				
Chen et al (2015)					x				
Jurdak et al (2015)	x	x	x		x				
Yin et al (2015)			x		x				
Blanford et al, (2015)	x	x							
Birkin et al (2014)		x			x				
Hawelka et al (2014)	x	x	x				x		
Li et al. (2013)								x	
Yuan et al (2013)			x						
Hasan et al (2013)			x						
Cranshaw et al. (2012)					x				
Nguyen e Szymanski (2012)		x						x	
Noulas et. al. (2011)		x	x						
Cheng et al (2011)	x	x	x						
Ferrari et al. (2011)					x				
Wakamiya et al. (2011)								x	x
Proposta apresentada	**x**	**x**			**x**				

Fonte: Produzido pelo autor

3.4 Sumarização das principais características dos trabalhos analisados

A revisão literária permitiu uma sumarização das principais características presentes nas pesquisas, facilitando a compreensão do estado da arte. A Tabela 3 exibe a sumarização dos referidos trabalhos considerando: (1) uso de dados sociais na pesquisa; (2) uso de dados governamentais abertos (e.g., dados de censos); (3) utilização de aspectos espaciais e temporais na pesquisa; (4) uso de dados relacionados ao círculo social dos usuários (e.g., quantidade de amigos na rede social); (5) utilização do conteúdo das mensagens na análise; (6) uso de múltiplas redes sociais no estudo; (7) Considera os POI visitados pelos usuários; (8) Considera o local de residência dos usuários.

Analisando a sumarização apresentada na Tabela 3, é possível perceber que todos os trabalhos presentes na literatura utilizam abordagens que consideram características espaciais e temporais dos dados, sendo esta característica inerente a este tipo de estudo. Porém, muitos trabalhos consideram apenas estes dois aspectos em seus estudos, não levando em conta, por exemplo, questões sociais e econômicas dos indivíduos.

Outra característica bastante observada nos trabalhos é a utilização do conteúdo das mensagens para a extração de dados semânticos. Nestes trabalhos, o conteúdo das mensagens é analisado para extrair, por meio de palavras ou termos utilizados (e.g “estou em casa”, “no trabalho”) informações que possam determinar, por exemplo, qual atividade o usuário está realizando, ou mesmo sua localização.

A Tabela 3 também deixa evidente que poucos trabalhos utilizam dados de cunho social, econômico ou demográfico nas suas pesquisas. Dos trabalhos levantados, apenas cinco utilizam algum tipo de indicador social extraído de dados governamentais. E, na maioria das vezes, esse uso é bastante superficial, não abrangendo um grande número de indicadores deste tipo, bem como realizando uma análise também superficial sobre seus possíveis relacionamentos com padrões de mobilidade.

Tabela 3 – Sumarização das principais características dos trabalhos apresentados

Trabalho	Considera aspectos sociais	Uso de dados governamentais abertos	Considera aspectos espaciais e temporais	Círculo Social	Utilização do conteúdo das mensagens	Uso de múltiplas redes sociais	Considera POI	Considera locais de residência
Gong (2016)	x		x		x	x	x	
Luo et al (2016)	x	x	x					x
Steiger et al. (2016)		x	x		x			
Steiger et al. (2015)	x	x	x		x			x
Chen et al (2015)			x			x		
Jurdak et al (2015)			x					
Yin et al (2015)			x		x	x	x	
Blanford et al, (2015)			x					
Birkin et al (2014)			x					x
Hawelka et al (2014)			x					
Li et al. (2013)	x	x	x			x		
Yuan et al (2013)			x		x			
Hasan et al (2013)			x		x	x		
Cranshaw et al. (2012)	x		x				x	
Nguyen e Szymanski (2012)			x	x				
Noulas et. al. (2011)			x					
Cheng et al (2011)	x	x	x		x			x
Ferrari et al. (2011)			x		x			
Wakamiya et al. (2011)			x					
Proposta apresentada	**x**	**x**	**x**				**x**	**x**

Fonte: Produzido pelo autor

3.5 Considerações finais

Neste capítulo, foi apresentado um levantamento dos trabalhos presentes na literatura que visam, fundamentalmente, extrair padrões de mobilidade de redes sociais e analisá-los, destacando suas características e, desta forma, enriquecer o conhecimento científico sobre este tipo de dado.

Por meio deste levantamento de trabalhos relacionados, foi possível observar a carência de estudos que utilizam aspectos sociais neste tipo de pesquisa, sendo isso evidenciado pelos próprios pesquisadores que, frequentemente, deixam explícito esta necessidade ao considerarem o uso de dados de cunho social para trabalhos futuros. Considerando este cenário, e somado ao fato de não ter sido encontrada na literatura nenhuma pesquisa que tivesse como objetivo a análise de correlação entre padrões de mobilidade e indicadores sociais, objetivando a descoberta de possíveis relacionamentos entre estas duas classes de dados, não é possível fazer uma comparação específica com os trabalhos relacionados e esta pesquisa.

Visando suprir a carência por trabalhos que apresentem uma análise mais aprofundada entre padrões de mobilidade e dados sociais, esta pesquisa (última linha - Tabela 2 e Tabela 3) propõe um método capaz de correlacionar padrões de mobilidade extraídos de redes sociais com indicadores sociais passados como parâmetro, permitindo assim, uma análise mais completa sobre a interação entre estas duas dimensões de dados, indo além da clássica abordagem espaço-temporal presente na literatura.

O próximo capítulo detalha as características do método proposto, bem como as técnicas utilizadas para o seu desenvolvimento.

Parte IV

Método para detecção e análise de padrões de mobilidade

4 Método para detecção e análise de padrões de mobilidade

Este capítulo tem como objetivo descrever o método proposto neste trabalho, o qual se destina a detectar padrões de mobilidade de usuários de redes sociais e correlacioná-los com indicadores sociais, econômicos e demográficos (aqui denominados apenas como indicadores sociais) presentes na região de estudo.

Para alcançar tal objetivo, foram desenvolvidos módulos que compõem o método proposto, os quais, trabalhando em conjunto, permitem gerar tabelas de correlação onde são expostas as informações relativas aos coeficientes de correlação entre as variáveis de mobilidade (padrões de mobilidade extraídos) e aos indicadores sociais fornecidos ao método em questão. Neste capítulo, todos estes módulos serão descritos em detalhes, apresentando-se seus papéis no funcionamento geral do método, bem como suas descrições formais por meio de pseudocódigos.

Como conjunto de dados para a extração de padrões de mobilidade, foram utilizadas mensagens georreferenciadas do Twitter, onde foram coletadas mensagens originadas na região da cidade de Londres durante o período de 26/11/2014 a 22/11/2015, totalizando 19.456.798 mensagens. É importante destacar que este estudo se aplica a qualquer outra rede social em que possa se obter mensagens contendo coordenadas geográficas.

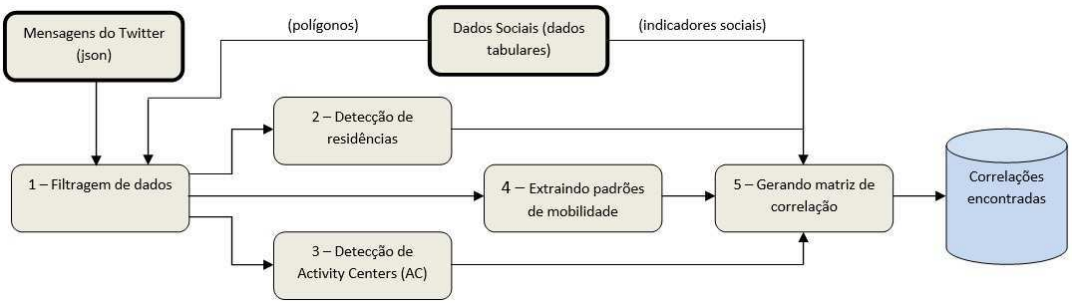
O restante do capítulo está organizado da seguinte forma. A seção 4.1 detalha as estruturas básicas do método proposto, descrevendo os dados utilizados na pesquisa, as técnicas utilizadas e as etapas de processamento implementadas. A Seção 4.2 descreve as considerações finais do capítulo.

4.1 Visão geral do método proposto

O método apresentado neste trabalho possui cinco diferentes etapas de processamento de dados onde, ao final, deseja-se obter uma matriz de correlação contendo os coeficientes de correlação entre todas as variáveis de mobilidade (padrões de mobilidade) extraídas das mensagens georreferenciadas e os indicadores sociais fornecidos ao método, permitindo assim, uma análise para determinar quais variáveis possuem maior grau de correlação entre si.

Para expressar de forma conceitual as principais etapas, exibe-se na Figura 8 um esquema demonstrando cada um dos passos do método, permitindo obter, ao final, a matriz de correlação.

Figura 8 – Fluxo de execução do método e geração da matriz de correlação



Fonte: Produzido pelo autor

O método apresentado recebe como entrada dois conjuntos de dados distintos. O primeiro consiste em mensagens georreferenciadas do Twitter, as quais serão utilizadas para minerar informações referentes aos usuários da rede, permitindo extrair informações referentes aos padrões de mobilidade dos mesmos, bem como permitirá a detecção das residências dos usuários, regiões frequentemente visitadas (AC) e também possíveis Pontos de Interesse (POI) que os mesmos costumam visitar em seus deslocamentos diários. O segundo conjunto de dados, que o método aceita como entrada, é referente aos indicadores sociais da região em que se deseja que o método execute suas análises, contendo, além dos indicadores sociais, as geometrias associadas a cada região da área em estudo, no caso deste trabalho, da cidade de Londres.

A partir dessas duas fontes de dados, o método apresentado é capaz de extrair os padrões de mobilidade para cada usuário presente nos dados, a partir de suas próprias mensagens postadas e, com isto, calcular as correlações destes padrões extraídos e os indicadores sociais relativos a área em estudo.

As seções seguintes deste capítulo visam apresentar em detalhes, tanto os dados que devem ser fornecidos ao método, quanto os módulos apresentados na Figura 8.

4.1.1 Mensagens do Twitter

Neste trabalho, cada mensagem do Twitter está em formato *json*, contendo todos os seus metadados disponibilizados pela API[1] da rede social. O conjunto de mensagens utilizado nesta pesquisa consiste em um subconjunto de *tweets* coletados por Oliveira (2017) referentes à região da cidade de Londres, permitindo assim, uma análise mais específica no contexto de uma grande metrópole europeia, a qual possui uma grande diversidade de pessoas, possuindo diferentes características, tanto no tocante a padrões de mobilidade quanto em aspectos sociais.

[1] Twitter API: <https://dev.twitter.com/>

É importante destacar que muitas destas mensagens coletadas não possuem coordenadas geográficas, sendo necessário um processo de filtragem de dados, permitindo a exclusão de mensagens que não possuam tais coordenadas. O precesso de filtragem será discutido em seções subsequentes deste trabalho.

4.1.2 Dados sociais

A escolha da cidade de Londres para o estudo de caso conduzido nesta pesquisa ocorreu não apenas pelo grande número de mensagens geradas por seus habitantes na rede social Twitter, mas também pela grande disponibilidade de dados e indicadores sociais disponibilizados por órgãos governamentais ligados à cidade. Para este estudo, foram utilizados diversos indicadores sociais, econômicos e demográficos referentes ao ano de 2011. Este ano foi escolhido por conter o maior número de indicadores, permitindo uma análise mais abrangente, maximizando o número de variáveis investigadas.

As variáveis utilizadas nesta pesquisa são referentes às seguintes categorias, totalizando 45 variáveis em estudo: (1) idade da população; (2) estrutura familiar; (3) grupos étnicos; (4) país de nascimento; (5) preços de imóveis de uma área; (nível de atividade econômica); (6) qualificação da força de trabalho; (7) nível de saúde da população; (8) disponibilidade de automóveis; (9) Religião.

O método aceita como entrada uma tabela onde cada coluna representa valores para um indicador social qualquer. Nesta tabela, uma destas colunas deve conter polígonos, podendo estes polígonos representarem, por exemplo, os bairros que compõem uma cidade. A Figura 9 apresenta um exemplo da tabela de indicadores sociais que podem ser utilizados.

Figura 9 – Organização estrutural dos indicadores sociais fornecidos ao método

other_religion numeric	no_religion numeric	geom geometry(MultiPolygon,4326)
3	567	0106000020E61000000100000
6	580	0106000020E61000000100000

Fonte: Produzido pelo autor

Neste trabalho, os indicadores sociais e mapas foram extraídos da plataforma governamental *London Datastore*, contendo dados do *Office for National Statistics*[2] (ONS). Esta plataforma reúne diversos dados relacionados à cidade de Londres, possuindo desde indicadores sociais, econômicos, culturais e políticos até mapas em formato *shapefile* e representações gráficas de toda a cidade de Londres, como pode ser visto na Figura 10.

[2] Contains National Statistics data © Crown copyright and database right [2012] and Contains Ordnance Survey data © Crown copyright and database right [2012].

A plataforma *London Datastore* demonstra ser uma ferramenta de extrema importância, tanto para a população desta cidade, quanto para o fomento de pesquisas científicas, facilitando o acesso aos indicadores sociais da região, bem como distribuindo livremente todo o seu acervo de dados.

Figura 10 – Mapa da cidade de Londres dividida em suas regiões distritais

Fonte: Produzido pelo autor

4.1.3 Filtragem de dados

O primeiro processo executado pelo método (etapa 1 – Figura 8) consiste na filtragem das mensagens do Twitter recebidas como entrada. Neste processo, são filtradas as mensagens que não possuem coordenadas geográficas em seus metadados, bem como aquelas em que suas coordenadas não apontam para dentro dos limites da área em estudo, no caso, da cidade de Londres. Esta filtragem é necessária pois a API do Twitter permite a busca por *tweets* a partir de um *bounding box* fornecido, de forma que muitas mensagens retornadas estão, na verdade, fora dos limites reais da cidade.

Dando continuidade ao processo de filtragem, mensagens postadas por usuários estacionários são removidas. Consideram-se usuários estacionários aqueles em que todas as mensagens postadas estão em um raio de 40 metros, representando, por exemplo, usuários que postam suas mensagens apenas de casa ou do trabalho. Este valor foi adotado como base para a detecção de agregações de mensagens, ou *clusters*, em todo este trabalho, onde mensagens com uma distância de até 40 metros entre si formariam uma agregação. Na literatura, existem trabalhos que utilizam diversos valores para este tipo de variável, como Montoliu, Blom e Gatica-Perez (2013) utilizando 250 metros e Kisilevich, Mansmann e Keim (2010) utilizando 30 metros. A partir de avaliação empírica, Neto, Baptista e

Campelo (2016) verificam que o valor de 40 metros se adéqua à detecção de *stay points* ou *clusters* semelhantes aos utilizados nesta pesquisa.

A filtragem supracitada é relevante devido ao fato de que muitos serviços postam mensagens originadas de uma mesma localização como, por exemplo, serviços de notícias, previsão de tempo, alertas de tráfego e demais serviços semelhantes. Também entram neste cenário usuários que postam mensagens exclusivamente de casa, por exemplo, sendo estes também desnecessários ao estudo. Finalmente, seguindo a metodologia proposta por Birkin et al. (2014), foram removidas todas as mensagens cujos usuários possuíam menos de 20 postagens, sendo estes desconsiderados devido à baixa atividade apresentada na rede, fato este que poderia levar a interferências severas nos resultados, não sendo estas mensagens capazes de representar, por exemplo, comportamentos recorrentes.

4.1.4 Detecção de residências

O estudo da região de residência é importante dado ao fato de que esta informação tende a expressar condições sociais e econômicas de um indivíduo, e estas condições podem, em parte, demonstrar como as pessoas se locomovem, especialmente em um ambiente urbano. Como exemplo, ao saber a região de residência de um indivíduo, bem como os indicadores sociais desta região, é possível verificar o nível de renda do local e traçar paralelos com outras áreas, onde a renda possivelmente será diferente, verificando como os padrões de mobilidade variam nestes casos.

Para a detecção das residências (etapa 2 - Figura 8), o método considera o local de maior intensidade de postagens durante a noite e início da manhã (LUO et al., 2016; HUANG; CAO; WANG, 2014) entre às 20 horas e 6 horas da manhã, bem como mensagens postadas durante a semana (segunda a sexta), consistindo este intervalo como sendo de maior permanência no local de residência. Para tal, o método proposto faz uso do algoritmo de agregação DBSCAN (*Density-based spatial clustering of applications with noise*) (ESTER et al., 1996). Este algoritmo tem como objetivo agregar pontos para formar *clusters* respeitando o raio mínimo entre um ponto e outro (ϵ); e o número mínimo de pontos que devem estar presentes em um *cluster* para ele existir (*minPts*). Este algoritmo tem como principais vantagens o fato de ser robusto frente à *outliers*, ou ruídos; pode encontrar *clusters* de tamanhos e formas variadas, bem como não necessita de informações prévias sobre os *clusters* presentes, utilizando apenas as duas variáveis citadas como parâmetro para a descoberta de *clusters* nos dados.

A partir da filtragem temporal descrita, executa-se o algoritmo de agregação supracitado sobre estas mensagens. O maior *cluster* encontrado é considerado como a região de residência do usuário, enquanto o ponto que representa o centro de massa deste *cluster* é considerado de fato como seu local de residência.

A Tabela 4 descreve o algoritmo de detecção de residências, onde este recebe uma lista com todos os usuários e calcula a localização estimada da residência para cada um deles.

No algoritmo supracitado, a linha 3 recupera as coordenadas de todas as mensagens que foram postadas durante os dias da semana e que estejam entre 20 horas e 6 horas da manhã. Estas coordenadas serão base para a detecção do local de residência para o usuário. Na linha 4, o DBSCAN é executado com o objetivo de agregar as mensagens que façam parte de *clusters* definidos por seus dois parâmetros. O resultado da execução deste método é uma lista de listas, onde cada elemento desta lista representa um *cluster* detectado pelo algoritmo. Já a linha 5 do código tem como objetivo identificar o maior *cluster* descoberto pelo DBSCAN. Este *cluster* será utilizado para calcular o centro de massa, estabelecendo assim, o local da residência em questão. Na linha 6 é calculado o centro de massa do maior *cluster* encontrado e, por fim, a última linha atribui as coordenadas do centro de massa como sendo o local de residência do usuário.

Tabela 4 – Código para a detecção de residências

1	function detectHomeLocatoin(listUsers)
2	for user in listUsers
3	listPoints = user.getAllMessagesAsPointsInHomeTime();
4	listOfClusters = executeDBSCAN(ϵ, minPts, listPoints);
5	biggestCluster = listOfClusters.getBiggestCluster();
6	centroidPoint = biggestCluster.calculateCentroid();
7	user.setHomePoint(centroidPoint);

Fonte: Produzido pelo autor

Com o objetivo de validar o método de detecção de residências utilizado, foram selecionados cinco voluntários para analisar imagens de satélite das regiões tidas como locais de residência. Cada voluntário analisou um conjunto de vinte imagens aleatórias (sem interseção entre os conjuntos) de pontos considerados como as residências dos usuários, objetivando assim, determinar, por meio dessas imagens, se os pontos estavam realmente em regiões residenciais, ou se estavam localizados em regiões que não condiziam com uma residência, em um bairro residencial. Os voluntários foram instruídos a classificar as imagens como: (1) região de residência, caso não houvesse quaisquer dúvidas quanto a esta classificação ou (2) indeterminado, caso não fosse possível identificar o ponto como estando relacionado a uma possível residência. A Figura 11 mostra um exemplo de imagem classificada como possível local de residência por um voluntário.

Como resultado da análise empregada, foram classificadas como locais de residência 63% das amostras analisadas.

Figura 11 – Exemplo de um centroide considerado como local de residência por um voluntário.

Fonte: Produzido pelo autor

4.1.5 Detecção de *Activity Centers*

Um AC pode ser definido como qualquer região em que um usuário frequentemente visita. A detecção de um AC mostra-se fundamental, especialmente por permitir uma análise mais detalhada de comportamentos rotineiros, permitindo, por exemplo, verificar em quais regiões da cidade os usuários costumam postar mais mensagens e como os indicadores sociais destas regiões se relacionam com os padrões de mobilidade dos indivíduos.

Para a detecção de AC, o método apresentado faz uso do algoritmo DBSCAN, o mesmo utilizado para a detecção das possíveis residências dos usuários. Porém, diferente da estratégia adotada na detecção de residências, aqui todos os *clusters* identificados pelo DBSCAN já são considerados como AC.

Para a análise e uso de AC neste trabalho, foram calculadas, para cada usuário presente nos dados, a mediana para todos os indicadores sociais associados às áreas geográficas onde seus *clusters* foram formados. Neste caso específico, a mediana foi adotada em detrimento da média pois, a partir de análises empíricas, foi verificado uma grande quantidade de *outliers* nos indicadores sociais, podendo estes virem a interferir nos resultados.

O método proposto para análise de AC permite estabelecer correlações estatísticas entre os valores das medianas de cada indicador social e os padrões de mobilidade extraídos para cada usuário. Isso permite, por exemplo, verificar se usuários com determinados

padrões de mobilidade tendem a visitar regiões onde a taxa de empregabilidade é maior.

A Tabela 5 demonstra o processo de detecção de AC onde, para cada usuário, é identificado todos os seus AC, bem como é calculado, para todos os AC encontrados, as medianas relativas aos indicadores sociais presentes nas regiões geográficas onde estes AC estão localizados.

A linha 2 itera sobre a lista de usuários passados para a função. A linha 3 recupera as coordenadas associadas a cada mensagem do usuário. A linha 4 executa o DBSCAN a partir de seus parâmetros e tendo como base as coordenadas recuperadas na linha anterior. O resultado da execução deste método é uma lista de listas, onde cada linha desta lista representa um *cluster* detectado pelo algoritmo.

Já nas linhas 5 e 6, os pontos de todos os *clusters* são adicionados em uma única lista. As linhas 7 e 8 iteram sobre a lista gerada na etapa anterior, buscando recuperar os indicadores sociais associados a cada região geográfica a qual cada ponto pertença. O método "*findSocialIndicatorsByRegion(point)*" retorna uma lista contendo os valores para cada indicador social associado à região onde o ponto passado como parâmetro pertença. Cada índice desta lista está associado a um determinado indicador social. A cada iteração, esta lista é adicionada a uma matriz, onde cada coluna representa os valores obtidos para cada indicador. A linha 9 calcula as medianas para cada coluna (indicador social) da matriz gerada na linha anterior, retornando uma lista com as medianas, onde cada elemento desta lista representa a mediana para cada indicador social presente na pesquisa. A última linha atribui a lista de medianas geradas na etapa anterior ao usuário.

Tabela 5 – Detectando Activity Centers e cálculo de medianas

1	function calculateMedianFromAC(listUsers)
2	for user in listUsers
3	listPoints = user.getAllMessagesAsPoints();
4	listOfClusters = executeDBSCAN(ϵ, minPts, listPoints);
5	for cluster in listOfClusters
6	listOfClusteredPoints.add(cluster.getAllPoints());
7	for point in listClusteredPoints
8	matrixSocialIndicators.add(findSocialIndicatorsByRegion(point));
9	listOfMedians = calculateMedians(matrixSocialIndicators);
10	user.setListOfMediansFromAC(listOfMedians);

Fonte: Produzido pelo autor

4.1.5.1 Detecção de Pontos de Interesse

Os POI em geral representam lugares (restaurantes, lojas, pontos turísticos entre outros) aos quais os indivíduos visitam em uma cidade. Assim como as regiões de residência, ou mesmo os AC descritos nas seções anteriores, os POI também trazem consigo informações relevantes a respeito das condições sociais e, principalmente, das condições econômicas de um indivíduo ou população.

Objetivando analisar como a visita frequente a diferentes POI, com diferentes preços pode estar associado a indicadores sociais, o método apresentando neste trabalho é capaz de extrair os POI a partir das mensagens postadas e correlacionar a faixa de preços adotadas pelos itens ou serviços prestados pelos estabelecimentos com os indicadores sociais fornecidos ao método. Para tal, são calculados os AC para cada usuário e extraído o centro de massa de cada AC. Obtendo este centro de massa, e sendo este um ponto, o método utiliza a API do Foursquare[3] para retornar um possível POI que esteja mais próximo a este ponto, bem como a classificação de preço associada ao POI, respeitando um raio máximo de 40 metros entre o centro de massa e o POI, de forma que, caso não exista nenhum POI dentro deste raio, nada é retornado. O valor de 40 metros foi obtido de forma empírica ao se analisar a ocorrência de um bom número de estabelecimentos dentro deste raio em relação aos centros de massa obtidos.

A Tabela 6 apresenta o processo de detecção de POI bem como a recuperação da classificação de preços, que varia de 1 (muito barato) a 4 (muito caro). Essa escala de preços é a mesma utilizada pela API do Foursquare.

Na Tabela 6, a linha 2 itera sobre a lista de usuários passada como parâmetro. Na linha 3, são recuperadas as coordenadas associadas a cada mensagem do usuário. Na linha 4, o DBSCAN é executado partir de seus parâmetros e tendo como base as mensagens recuperadas na linha anterior. O resultado da execução deste método é uma lista de listas, onde cada elemento desta lista representa um *cluster* detectado pelo algoritmo. A linha 5 do algoritmo itera sobre a lista gerada na linha anterior. Na linha 6 é calculado o centro de massa para cada *cluster* identificado pelo DBSCAN. A linha 7 verifica, através da API do Foursquare, se o centro de massa calculado está relacionado a algum POI existente. Caso exista algum POI dentro do raio de 40 metros, a função retorna o POI mais próximo ao centro de massa, bem como retorna todas as informações disponíveis sobre o local, incluindo a classificação de preço associada. As linhas 8 e 9 verificam se algum POI foi retornado na linha anterior e, em caso afirmativo, o POI é vinculado ao usuário.

Tabela 6 – Detectando pontos de interesse com auxílio da API do Foursquare

1	function detectPointOfInterest(listUsers)
2	for user in listUsers
3	listPoints = user.getAllMessagesAsPoints();
4	listOfClusters = executeDBSCAN(ϵ, minPts, listPoints);
5	for cluster in listOfClusters
6	centroid = calculateCentroid(cluster);
7	poi = findFoursquarePOI(centroid);
8	if(poi != null)
9	user.getPOIList().add(poi);

Fonte: Produzido pelo autor

[3] Foursquare: <https://developer.foursquare.com/>

4.1.6 Extraindo padrões de mobilidade

Esta etapa (etapa 4 - Figura 8) consiste na extração das propriedades estatísticas que descrevem os padrões de mobilidade. Neste estudo, além da proposição de novas métricas de mobilidade, serão utilizadas duas métricas largamente utilizadas em pesquisas abrangendo o escopo de mobilidade, sendo elas o Raio de Giro e a distância total percorrida (LUO et al., 2016; CHENG et al., 2011; GONZALEZ; HIDALGO; BARABASI, 2008; HASAN; ZHAN; UKKUSURI, 2013).

Também são propostos neste trabalho filtros temporais específicos que visam, além de refinar a análise de padrões de mobilidade, permitir a descoberta de características que possam ser observadas apenas em ocasiões específicas como, por exemplo, em finais de semana ou em feriados.

As seções seguintes visam detalhar as variáveis de mobilidade utilizadas para a captura de padrões de mobilidade presentes em usuários de redes sociais.

4.1.6.1 Raio de Giro

O Raio de Giro corresponde ao desvio padrão de distâncias entre os pontos que representam os deslocamentos e o centro de massa destes pontos. Esta métrica permite avaliar quão longe um indivíduo se desloca e quão frequentes são estes deslocamentos.

O Raio de Giro representa uma estatística importante no que concerne aos padrões de mobilidade, pois, através de um único valor, pode-se determinar características determinantes para o estudo de mobilidade. Por exemplo, um baixo valor de Raio de Giro significa que o indivíduo tende a se locomover por curtas distâncias, com poucos deslocamentos mais longos. Já um Raio de Giro alto, tende a expressar características de um indivíduo que se desloca frequentemente por longas distâncias. A Equação 4.1 representa a formalização desta métrica:

$$r = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(p_i - p_c)^2} \tag{4.1}$$

Para a equação, temos que:

a) r representa o valor para o Raio de Giro de um indivíduo;

b) m representa o número de mensagens de um indivíduo;

c) p_i expressa um ponto onde uma mensagem foi postada;

d) p_c representa o centro de massa das mensagens de um indivíduo;

e) $(p_i - p_c)$ é a distância entre um ponto de uma mensagem e seu centro de massa;

4.1.6.2 Distância Total Percorrida

A distância total percorrida representa a soma das distâncias entre todos os deslocamentos consecutivos realizados pelo usuário, refletindo a distância total percorrida deste usuário dentro da área de estudo.

Diversos autores adotam esta métrica em seus estudos sobre padrões de mobilidade. Cheng et al. (2011) sugerem que o comportamento desta métrica para mensagens coletadas do Twitter tende a seguir a distribuição de Lévy Flight, a qual é caracterizada por deslocamentos curtos e aleatórios, com eventuais deslocamentos longos. Shin et al. (2008) encontram resultados semelhantes na distribuição desta métrica ao analisar dados oriundos de dispositivos GPS usados em diferentes cenários, tais como em áreas metropolitanas e em um campus estudantil.

Opondo-se aos estudos que apontam esta tendência, Gonzalez, Hidalgo e Barabasi (2008) analisam dados coletados de redes de celulares e verificam que os deslocamentos humanos possuem um grau significativo de regularidade espacial e temporal. Isto deve-se, principalmente, ao fato de indivíduos tenderem a retornar a lugares que já tinham visitado anteriormente, sendo observado também que os últimos locais visitados possuem mais chances de serem visitados novamente, revelando assim uma relação temporal referente à probabilidades de retorno.

As próximas quatro subseções descrevem as novas métricas de mobilidade propostas neste trabalho.

4.1.6.3 Número de Deslocamentos

Com o objetivo de captar comportamentos que não poderiam ser percebidos utilizando-se as duas métricas de mobilidade apresentadas inicialmente (Raio de Giro e Distância Total Percorrida), esta pesquisa propõe quatro novas métricas de mobilidade. A primeira nova métrica, denominada de Número de Deslocamentos, visa determinar a quantidade de deslocamentos realizados entre diferentes lugares, ou seja, o número de movimentações que caracterizaram alguma mudança real de local pelo usuário. Por exemplo, para um usuário que se desloque em um determinado dia, de sua residência até o local de trabalho e depois retorne para sua residência, contabilizam-se apenas duas movimentações entre lugares, uma de casa para o trabalho, e outra do trabalho para casa, desde que o local da residência esteja a uma distância superior a 40 metros. Mensagens postadas dentro de um raio de 40 metros não são contabilizadas como um deslocamento válido entre diferentes lugares. A adoção deste valor está justificada na seção 3.1.3 deste capítulo.

A Figura 12 demonstra um exemplo de deslocamento entre diferentes lugares, onde um usuário posta diversas mensagens perto da *London Eye* dentro de um raio de 40 metros

e, em seguida, seguindo um ordenamento temporal, desloca-se para o *Sea Life London Aquarium*, onde volta a postar novas mensagens, totalizando um único deslocamento entre lugares.

Figura 12 – Deslocamento entre diferentes lugares adotados no método

Fonte: Produzido pelo autor utilizando a ferramenta *Scribble Maps*

4.1.6.4 Média de Deslocamentos Por Dia

Esta variável consiste em analisar o número de deslocamentos entre diferentes lugares (Seção 4.1.6.3) porém considerando a média destes deslocamentos realizados por dia, e apenas os deslocamentos que se iniciem e terminem no mesmo dia.

Esta variável de mobilidade permite capturar comportamentos associados a rotinas de trabalho, dado que estes, em geral, ocorrem em dias úteis e possuem características bem definidas, como tempo de duração e distância entre casa e trabalho.

A Média de Deslocamentos Por Dia permite, por exemplo, verificar possíveis correlações entre o número de deslocamentos diários e indicadores sociais como, por exemplo a renda ou mesmo o nível de qualificação dos indivíduos, o que, junto com outros indicadores apresentados neste trabalho, poderiam ajudar a verificar se, por exemplo, pessoas mais ricas tendem, ou não, a se deslocarem um maior número de vezes por dia.

4.1.6.5 Média de Distância Percorrida Por Deslocamentos

Assim como a variável descrita na seção anterior, a Média de Distância Percorrida Por Deslocamentos analisa o número de deslocamentos entre diferentes lugares (Seção 4.1.6.3) porém agora considerando as distâncias percorridas nestes deslocamentos.

Esta nova variável de mobilidade permite analisar padrões associados às distâncias entre deslocamentos entre diferentes regiões e, em conjunto com a métrica anterior, permite verificar, por exemplo, características de usuários que moram mais perto ou mais distantes dos seus locais de trabalho, condição esta que, utilizando apenas as métricas de Raio de Giro e Distância Total Percorrida, não seria possível identificar.

Esta variável permite, além de complementar os indicadores de mobilidade apresentados nas Seções 4.1.6.3 e 4.1.6.4, verificar como a distância entre estes deslocamentos se correlacionam com os indicadores sociais e, trazendo o exemplo da seção anterior, saber se, pessoas mais ricas tendem, ou não, a se deslocarem não somente um maior número de vezes, mas também por maiores distâncias.

4.1.6.6 Média de Preços de POI Visitados

Nesta pesquisa, este indicador está relacionado à faixa de preços utilizados pelos POI frequentemente visitados pelos usuários, permitindo analisar como a frequência de visitas a POI com preços mais elevados ou mais baixos está associada a determinados indicadores sociais destes usuários.

Para esta análise, assim como demonstrado na Seção 4.1.5.1, onde se descreve a extração dos POI utilizados, são considerados apenas os locais que possuem algum valor associado ao preço adotado no estabelecimento. Quaisquer outros locais onde a informação relativa a preços não esteja disponível, são desconsiderados da análise.

A faixa de preços utilizada nesta pesquisa também é extraída a partir da API do Foursquare, variando de 1 (muito barato) a 4 (muito caro). Os dados fornecidos pela API do Foursquare possuem a vantagem de serem gerados pela própria comunidade de usuários desta rede, sendo esta informação atualizada constantemente e, portanto, refletindo as reais características de cada estabelecimento frequentado.

Como exemplo prático, esta variável permite verificar se usuários que costumam visitar POI com preços mais elevados tendem a residir ou mesmo frequentar regiões onde o número de imigrantes é maior ou menor, podendo demonstrar que este indicador social poderia estar vinculado a condições de renda da população.

É importante destacar que, devido ao fato de terem sido considerados apenas POI que tivessem preços associados a estes, locais como praças, pontos turísticos e demais localidades onde a qualificação associada ao preço não se aplica, não puderam ser analisados pelo método proposto.

4.1.7 Gerando a matriz de correlação

A última etapa exibida pelo fluxo de execução (etapa 5 - Figura 8) tem por objetivo calcular as correlações entre os padrões de mobilidade extraídos em estágios anteriores

e os indicadores sociais recebidos como entrada, gerando uma matriz em um arquivo *.xls* contendo todos os coeficientes de correlação entre estas duas classes de variáveis. A Tabela 7 mostra um fragmento desta matriz de correlação. Os valores em azul representam coeficientes de correlação considerados como relevantes.

Tabela 7 – Fragmento da matriz de correlação gerada pelo método

Variáveis de mobilidade / variáveis sociais	Total de pessoas economicamente inativas	Total de pessoas economicamente ativas desempregadas	Taxa de empregabilidade	Pessoal sem qualificações profissionais
Raio de Giro	-0.3002	-0.1354	0.2698	-0.2684
Distância Total Percorrida	-0.2462	-0.2693	0.1936	-0.3002
Número de Deslocamentos	-0.0667	-0.0287	0.0015	-0.0731
Média de Deslocamentos Por Dia	0.0651	0.0717	-0.0920	0.0428
Média de Distância Percorrida Por Deslocamentos	-0.2494	-0.2374	0.2317	-0.2112
Média de preços de POI	-0.1729	-0.0946	0.1055	-0.2493

Fonte: Produzido pelo autor

Para o cálculo das correlações, o método permite que sejam executados os três testes mais utilizados na literatura, que são: Pearson, Spearman e Kendall (CHOK, 2010). A escolha do teste a ser executado é uma atividade que depende de um conhecimento prévio acerca dos dados que serão analisados no estudo, devendo-se levar em consideração diversos aspectos sobre estes dados como, por exemplo, uma possível não-normalidade nestes, a incidência de empates, linearidade no relacionamento entre as variáveis e diversos outros aspectos importantes para uma execução precisa dos testes de correlação. Detalhes sobre a escolha do teste utilizado, bem como demais considerações estão mais adiante no capítulo que trada dos experimentos executados.

4.2 Considerações finais

Neste capítulo, foi apresentado um método computacional capaz de extrair padrões de mobilidade de usuários de redes sociais e permitir a correlação destes padrões com indicadores sociais desta população.

Este capítulo mostrou todo o fluxo de processamento de dados através dos diferentes módulos apresentados na Figura 8. Adicionalmente, foram apresentados aspectos

relacionados à extração de padrões de mobilidade e técnicas específicas de agregação (*clustering*) empregadas no trabalho, com o objetivo de auxiliar na detecção de residências, AC e POI frequentados pelos usuários da rede.

No tocante à extração de padrões de mobilidade, diversos desafios foram encontrados, especialmente no que se refere à natureza esparsa e descontínua das mensagens postadas no Twitter, tornando a análise deste tipo de dado mais difícil e imprecisa. Estes aspectos serão discutidos com mais detalhes no capítulo de experimentos.

Com todas as características discutidas neste capítulo, o método apresentado se diferencia de demais técnicas presentes na literatura por oferecer uma abordagem automatizada de extração e análise de padrões de mobilidade e suas possíveis correlações com indicadores sociais. O método retorna como saída uma matriz de correlação, possibilitando a identificação das correlações mais relevantes, o que pode, além de criar subsídios ao melhor entendimento da dinâmica urbana, auxiliando, por exemplo, a tomada de decisões de gestores públicos, vir a fornecer informações úteis a sistemas de recomendação de lugares.

O próximo capítulo apresenta os experimentos conduzidos para validar o método aqui apresentado.

Parte V

Avaliação Experimental

5 Avaliação experimental

Neste capítulo, serão apresentados os experimentos conduzidos com o intuito de validar o método proposto nesta dissertação. Estes experimentos são organizados em dois grupos distintos, referenciados como Experimento 1 e Experimento 2, e visam verificar a correlação de padrões de mobilidade com indicadores sociais. No Experimento 1, verifica-se como os padrões de mobilidade dos indivíduos estão relacionados com a região de suas residências; O Experimento 2, por sua vez, tem como objetivo verificar como os padrões de mobilidade dos indivíduos estão relacionados com os indicadores sociais das regiões frequentemente visitadas por estes.

As demais seções deste capítulo estão organizadas da seguinte forma: a Seção 5.1 descreve o conjunto de dados utilizados no trabalho; a Seção 5.2 descreve os experimentos realizados na pesquisa; a Seção 5.3 descreve os resultados para o primeiro experimento, que relaciona padrões de mobilidade de indivíduos com os indicadores sociais da região de suas residências; a Seção 5.4 descreve os resultados para o segundo experimento, que relaciona padrões de mobilidade de indivíduos com os indicadores sociais das regiões frequentemente visitadas (AC) por estes; a Seção 5.5 discute as principais limitações encontradas nos resultados obtidos; a Seção 5.6 discute em conjunto os resultados dos dois experimentos realizados. Finalmente, a Seção 5.7 encerra o capítulo apresentando as considerações finais.

5.1 Conjunto de dados

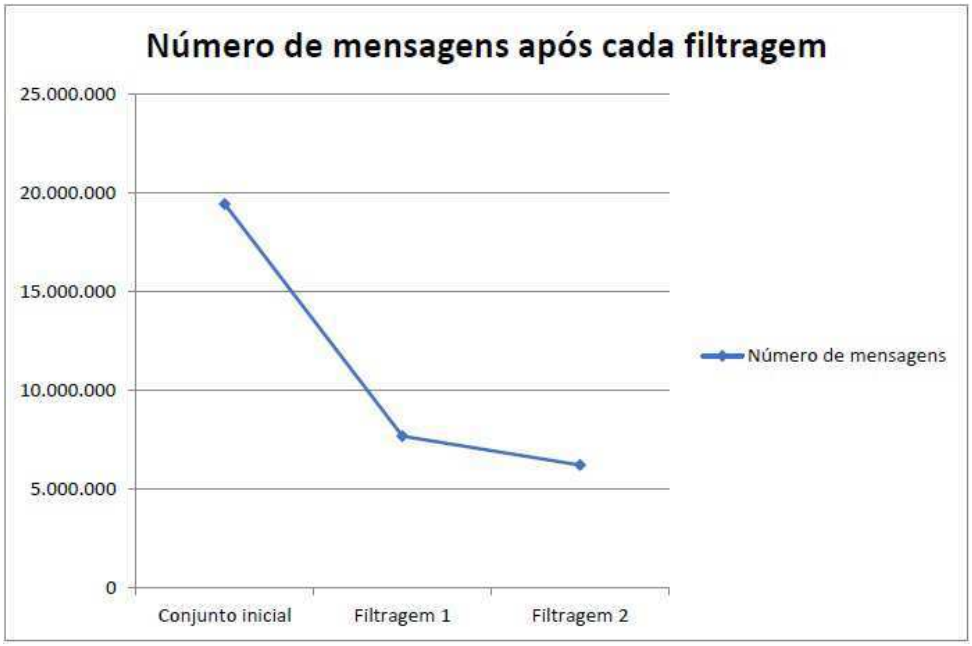
O conjunto de dados utilizado nesta pesquisa compreende mensagens do Twitter e indicadores sociais da região em estudo. Das 19.456.798 mensagens coletadas do Twitter, foi extraído um total de 568.322 usuários da rede social.

Partindo do conjunto inicial dos dados, o método considera apenas mensagens que possuam coordenadas geográficas (latitude e longitude), excluindo as que não possuem este metadado (filtragem 1). Com essa exclusão, o conjunto de mensagens se reduziu a 7.680.200 mensagens com as referidas coordenadas, compreendendo um total de 351.656 usuários. Após essa filtragem, foram removidas as mensagens que não estavam localizadas dentro dos limites geográficos da cidade de Londres, bem como mensagens postadas por usuários estacionários e por usuários com menos de 20 mensagens postadas, reduzindo assim o conjunto a 6.215.792 mensagens e 53.093 usuários (filtragem 2), com uma média de 117,07 mensagens por usuário.

A Figura 13 exibe a evolução no número de mensagens ao longo do processo de filtragem e a Figura 14 exibe esta mesma evolução, porém relacionado ao número total de

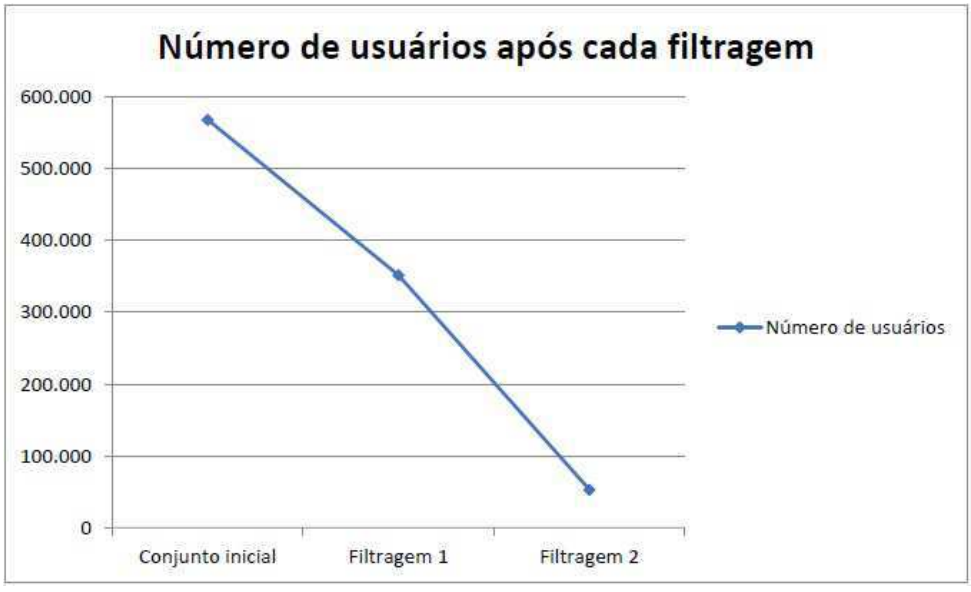
usuários, deixando evidente a redução em ambos os conjuntos de dados quando submetidos a cada etapa de filtragem.

Figura 13 – Gráfico de evolução das mensagens ao longo das etapas de filtragem



Fonte: Produzido pelo autor

Figura 14 – Gráfico de evolução do número total de usuários ao longo das etapas de filtragem

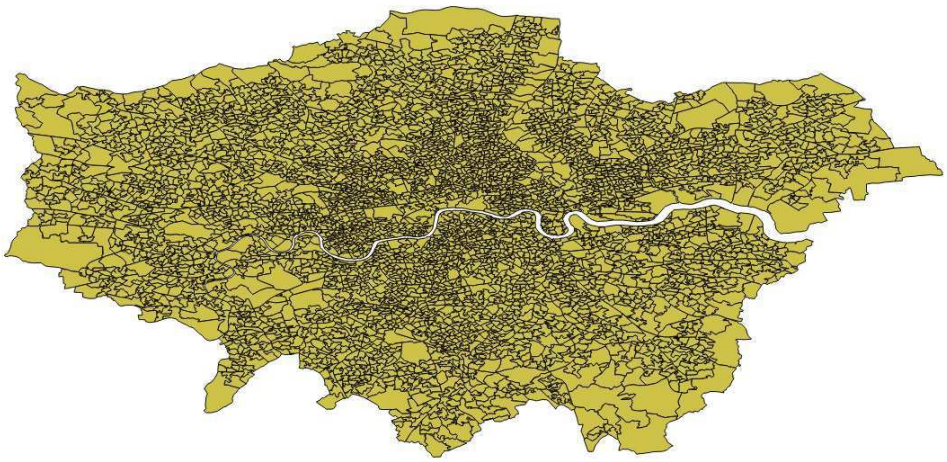


Fonte: Produzido pelo autor

Os dados sociais foram coletados da plataforma *London Datastore*, plataforma aberta ao público que contempla dados de censos demográficos bem como diversas classes de indicadores sociais, econômicos e demográficos da cidade de Londres. Os dados coletados foram referentes ao ano de 2011, contemplando as seguintes classes de indicadores: (1) Faixa etária da população; (2) Estrutura familiar; (3) Grupos étnicos; (4) País de nascimento; (5) Preço de imóveis; (6) Atividade econômica; (7) Qualificação profissional; (8) Níveis de saúde; (9) Disponibilidade de automóveis por família; (10) Religião.

Para esta pesquisa, a região de Londres foi dividida utilizando-se o conceito de LSOA[1] (*Lower Super Output Area*), que corresponde a menor subdivisão da área da cidade de Londres disponível nos dados colhidos. Nesta subdivisão, cada área possui em média 1.722 habitantes. A Figura 15 mostra o mapa da cidade de Londres utilizado na pesquisa subdividido em LSOA.

Figura 15 – Mapa da cidade de Londres subdividido em LSOA



Fonte: Produzido pelo autor

5.2 Design de experimentos

Os experimentos conduzidos neste trabalho visam responder às seguintes questões de pesquisa:

a) Questão de pesquisa (Q1): É possível estabelecer correlações estatísticas significativas entre padrões de mobilidade e dados sociais?

[1] LSOA atlas: <https://data.london.gov.uk/dataset/lsoa-atlas>

  - Hipótese nula (H0 - Q1): Não é possível estabelecer correlações estatísticas significativas entre padrões de mobilidade e dados sociais.
  - Hipótese alternativa (H1 - Q1): É possível estabelecer correlações estatísticas significativas entre padrões de mobilidade e dados sociais.

b) Questão de pesquisa (Q2): Existe correlação entre padrões de mobilidade de um indivíduo e os indicadores sociais da região de sua residência?
  - Hipótese nula (H0 - Q2): Não existe correlação entre padrões de mobilidade de um indivíduo e os indicadores sociais da região de sua residência.
  - Hipótese alternativa (H1 - Q2): Existe correlação entre padrões de mobilidade de um indivíduo e os indicadores sociais da região de sua residência.

c) Questão de pesquisa (Q3): Existe correlação entre os padrões de mobilidade de um indivíduo e os indicadores sociais presentes em seus AC?
  - Hipótese nula (H0 - Q3): Não existe correlação entre os padrões de mobilidade de um indivíduo e os indicadores sociais presentes em seus AC.
  - Hipótese alternativa (H1 - Q3): Existe correlação entre os padrões de mobilidade de um indivíduo e os indicadores sociais presentes em seus AC.

No tocante às questões de pesquisa citadas, a questão Q1 visa, fundamentalmente, definir de forma geral a viabilidade do estudo proposto, demonstrando se foram ou não encontrados coeficientes de correlação significativos para o estudo em questão, abrindo assim portas para a verificação das demais questões de pesquisa.

A questão de pesquisa Q2 tem como objetivo verificar possíveis correlações entre os padrões de mobilidade extraídos para um indivíduo e os indicadores sociais presentes na região de sua residência, permitindo assim analisar quais variáveis estariam mais relacionadas entre si, bem como a magnitude e direção da possível correlação. Esta questão permitirá verificar, por exemplo, se um usuário que mora em uma região mais rica tende a se movimentar por maiores distâncias.

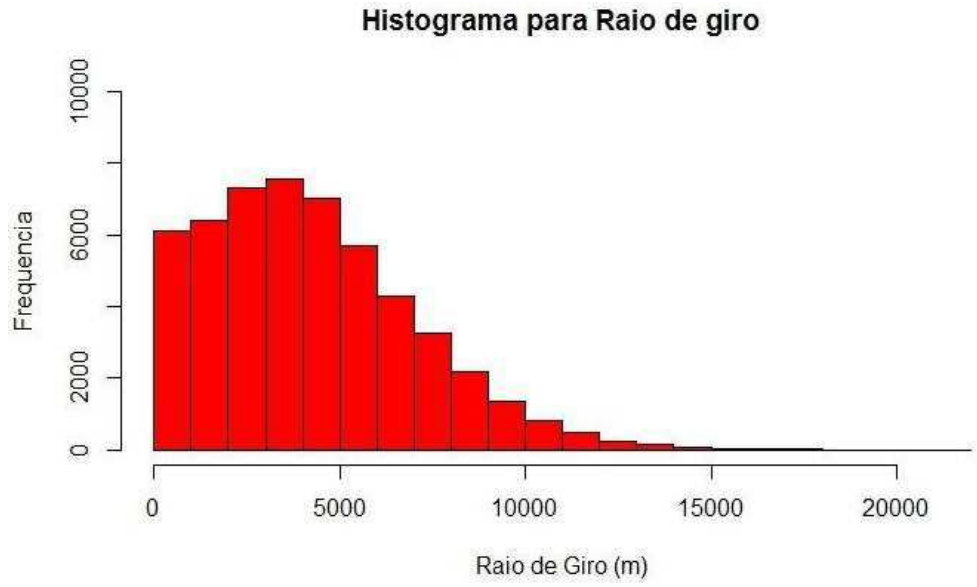
A questão Q3 visa estudar possíveis correlações entre os padrões de mobilidade de um indivíduo e as regiões frequentemente visitadas por ele, tratadas neste trabalho como os AC. Esta questão permite, por exemplo, verificar se indivíduos com um Raio de Giro maior tendem a frequentar locais onde os preços dos imóveis são maiores.

5.2.1 Configurações gerais dos experimentos

Após as filtragens iniciais executadas pelo método, as quais são descritas no Capítulo 4, são extraídos todos os padrões de mobilidade para os usuários que satisfazem as condições estabelecidas na etapa de pré-processamento. As Figuras de 16 a 21 demonstram a organização dos padrões de mobilidade extraídos. Para as Figuras de 17 a 20, os valores são exibidos em escala de log10.

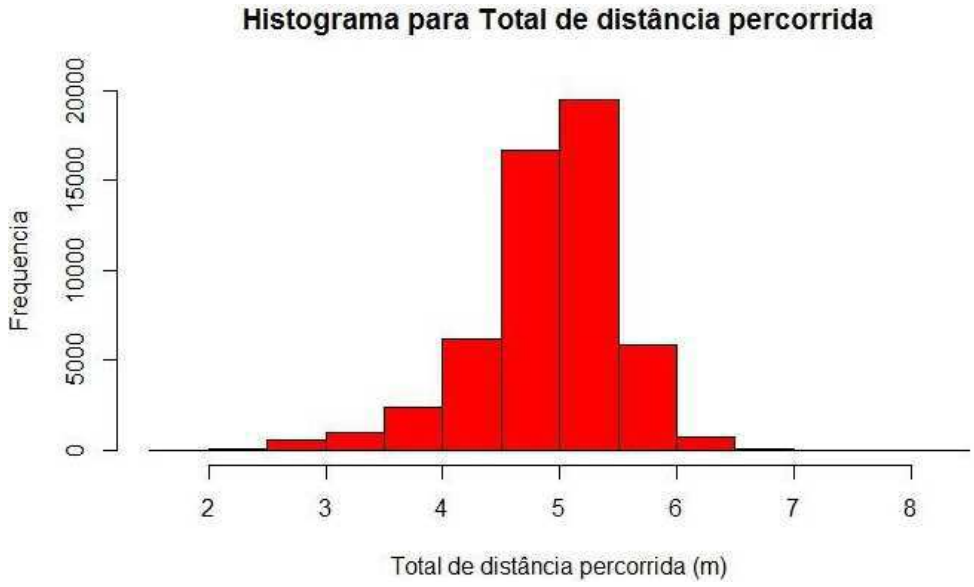
Para a Figura 16, que exibe os dados para a variável de mobilidade Raio de Giro, é possível verificar que os valores para esta variável estão, em sua maioria, entre 3.000 e 4.000 metros aproximadamente, totalizando 7.554 usuários nesta faixa de valores.

Figura 16 – Histograma para a variável de mobilidade Raio de Giro



Fonte: Produzido pelo autor

Figura 17 – Histograma para a variável de mobilidade Total de Distância Percorrida (log 10)



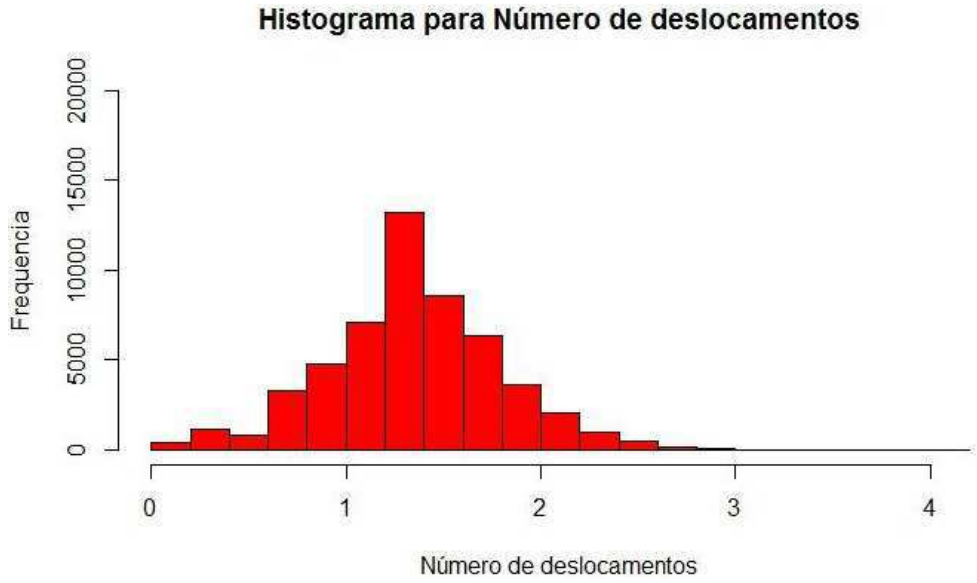
Fonte: Produzido pelo autor

Para a Figura 17, pode-se perceber que grande parte dos seus valores estão situados

aproximadamente entre 100.000 metros e 316.228 metros, compreendendo um número de usuários dentro deste intervalo de 19.514 usuários.

Para a Figura 18, que trata dos valores obtidos para a variável de mobilidade "Número de Deslocamentos", foram observados picos de valores entre 15 a 25 deslocamentos por usuários, aproximadamente. Esta faixa de valores compreende o total de 14.710 usuários.

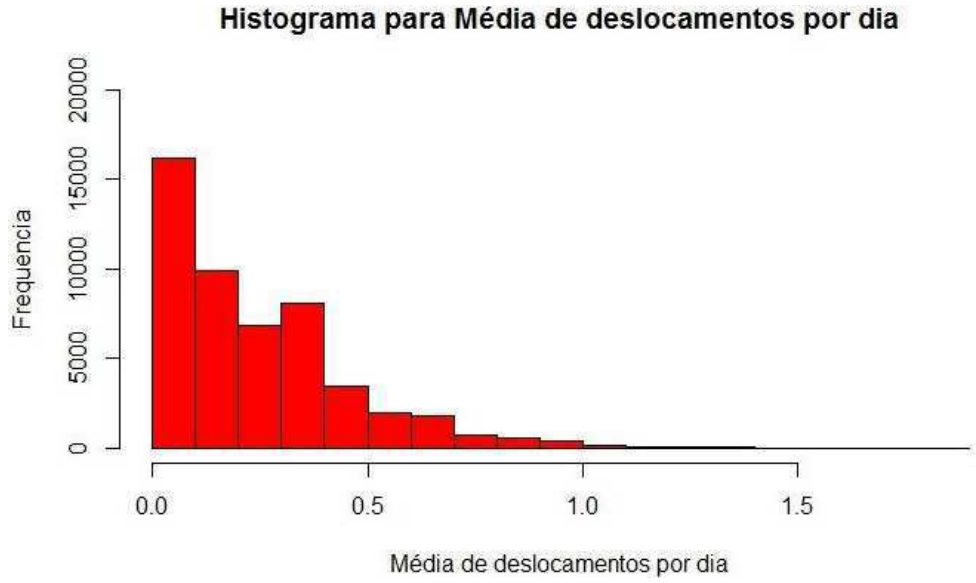
Figura 18 – Histograma para a variável de mobilidade Número de Deslocamentos (log 10)



Fonte: Produzido pelo autor

A Figura 19 exibe os valores para a variável "Média de Deslocamentos Por Dia", onde a maior incidência de valores situa-se entre 0 e 1,2 deslocamentos médios por dia, com um total de 16.900 usuários. Usuários com valores iguais a zero para esta variável podem ser explicados pelo fato de que, para estes, não foram observados deslocamentos superiores à 40 metros em um mesmo dia, sendo estas mensagens postadas sempre de um mesmo ponto, dentro de um intervalo de um dia.

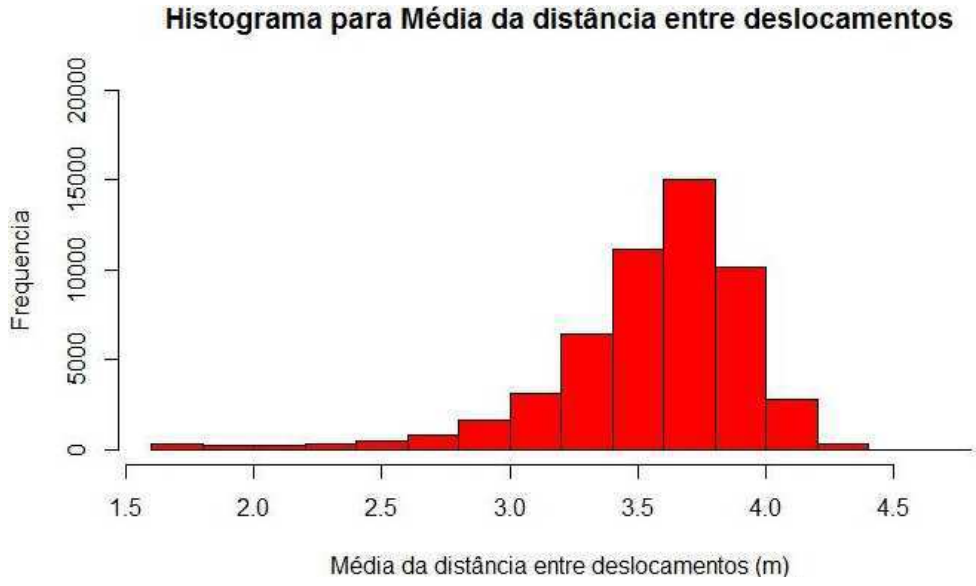
Figura 19 – Histograma para a variável de mobilidade Média de Deslocamentos Por Dia (log 10)



Fonte: Produzido pelo autor

Para a variável "Média de Distância Entre Deslocamentos", a Figura 20 projeta a distribuição desta variável nos dados, onde a maioria dos usuários possuem valores entre 3.981 e 6.309 metros aproximadamente, totalizando 15.016 usuários.

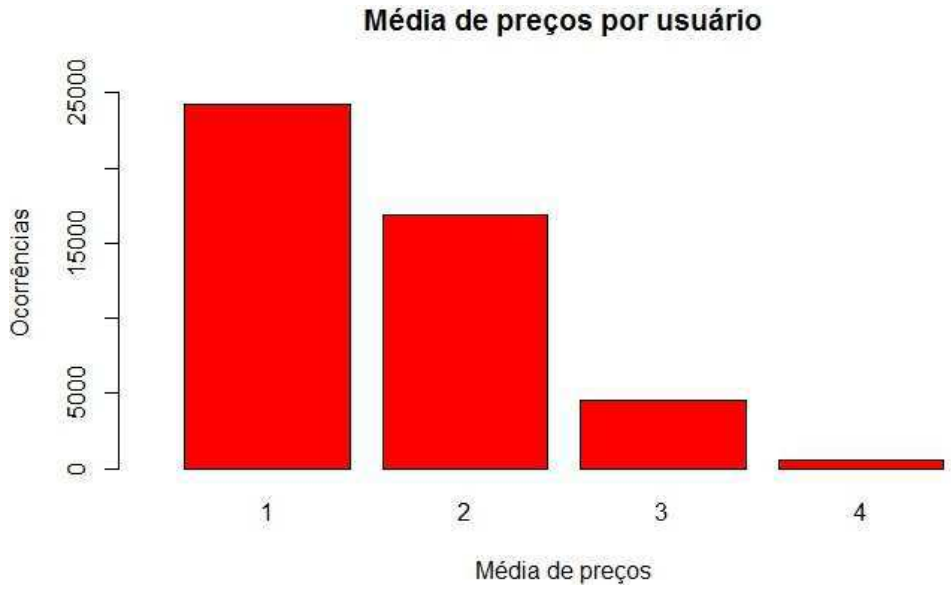
Figura 20 – Histograma para a variável de mobilidade Média da distância entre deslocamentos (log 10)



Fonte: Produzido pelo autor

A Figura 21 exibe os valores associados às médias de preços de POI frequentemente visitados pelos usuários da rede social. Os valores são distribuídos em uma escala de 1 a 4, onde o menor valor está relacionado a POI com preços mais baixos e o valor maior está relacionado a POI com preços mais elevados. Para esta variável, é possível visualizar o grande número de visitas a POI com preços mais baixos (valor 1), se comparado a outras faixas de preço.

Figura 21 – Gráfico em barras para a variável Média de Preços de POI Visitados



Fonte: Produzido pelo autor

Para a execução dos experimentos propostos neste trabalho, os usuários que tiveram suas mensagens coletadas e filtradas foram agrupados nas seguintes categorias:

a) Categoria 1: Usuários com pelo menos 1.000 mensagens postadas (635 usuários);

b) Categoria 2: Usuários com pelo menos 2.500 mensagens postadas (153 usuários);

c) Categoria 3: Usuários com pelo menos 5.000 mensagens postadas (36 usuários);

Esta divisão foi estabelecida com o objetivo de identificar possíveis correlações que estivessem presentes apenas em usuários que fizessem um uso intenso da rede social, os quais teriam um maior número de mensagens postadas, facilitando assim a extração dos padrões de mobilidade.

Também como característica comum a todos os experimentos aqui propostos, cita-se a utilização do algoritmo DBSCAN para agregar pontos, tanto para a detecção de residências quanto para a detecção de AC. Em linhas gerais, o DBSCAN irá considerar pontos agregados em um *cluster* em uma distância máxima de ϵ metros, bem como serão necessários ao menos *minPts* pontos para a formação de um *cluster*. Pontos que não

se enquadrem nestes parâmetros serão tratados como ruídos pelo algoritmo. Para os experimentos, o algoritmo foi configurado com os seguintes valores:

a) Detecção de residências: $\epsilon = 40$m; $minPts = 4$;

b) Detecção de AC: $\epsilon = 40$m; $minPts = 3$;

c) Detecção de POI com preços disponíveis: $\epsilon = 40$m; $minPts = 3$.

Para os testes de correlação, tanto para o experimento relacionado à questão Q2 quanto para o experimento relacionado à questão Q3, utiliza-se o teste de correlação de Kendall *tau-b*, tendo seu coeficiente de correlação representado por τ. Este teste se caracteriza como um teste de correlação não-paramétrico, sendo adequando à condição de não-normalidade dos dados utilizados na pesquisa, sendo também mais resistente à presença de dados repetidos, permitindo uma análise mais fidedigna no tocante às correlações.

Ao compararmos o teste de correlação de Kendall com seu concorrente não-paramétrico, o teste de Spearman, o último tendeu a apresentar coeficientes de correlações mais altos que os demonstrados pelo teste de Kendall, possivelmente pelo fato de o teste de Spearman não tratar de forma eficaz os dados repetidos presentes no conjunto. Estes dados podem ser observados, quando, por exemplo, dois usuários postam mensagens em uma mesma localização, de forma que seus indicadores sociais seriam os mesmos nesta situação.

Ainda no tocante à análise de correlações, este trabalho considera apenas as correlações onde $\tau >= 0{,}25$, com uma significância estatística onde $p\text{-}value < 0{,}05$. Cohen (1988 apud MILES; SHEVLIN, 2001) definiu escalas para valores de correlação, onde valores de aproximadamente 0,1 seriam correlações fracas, correlações médias teriam valores de aproximadamente 0,3 e correlações altas seriam maiores ou iguais a 0,5.

Objetivando encontrar correlações em determinados períodos ou momentos da vida cotidiana, também foram elaborados junto ao método, níveis de filtragem que permitem, por exemplo, executar os testes em mensagens geradas apenas em feriados, ou mesmo nos finais de semana. Esta metodologia permite aplicar um olhar mais detalhado no que tange os hábitos de mobilidade das pessoas que se locomovem dentro do espaço urbano estudado.

Além de considerar mensagens sem nenhum tipo de filtro temporal, este trabalho divide as mensagens de acordo com os seguintes filtros temporais:

a) mensagens postadas apenas em feriados (*bank holidays*);

b) mensagens postadas apenas em dias úteis;

c) mensagens postadas apenas durante os finais de semana (sábado e domingo);

d) mensagens postadas durante feriados e aos domingos;

5.3 Experimento 1: análise de correlação entre padrões de mobilidade e o local de residência

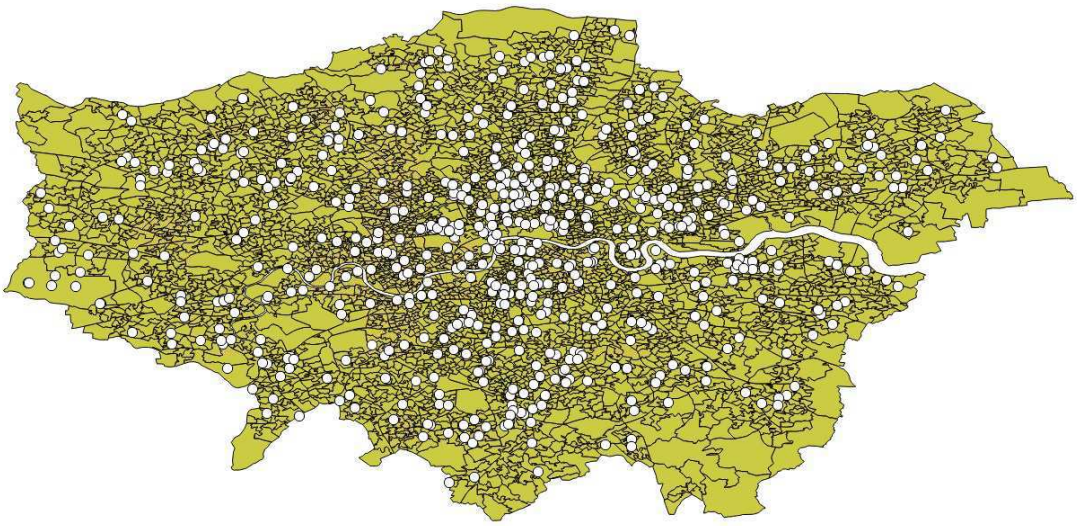
Este primeiro experimento visa avaliar a correlação entre o local de residência de um indivíduo e seus padrões de mobilidade extraídos pelo método. O local de residência foi escolhido devido ao contexto social que esta informação traz, principalmente em questões como renda e qualidade de vida dos habitantes.

Para a detecção das residências, são agregadas as mensagens postadas por todos os usuários entre os horários de 8pm a 6am, durante os dias da semana (segunda-feira a sexta-feira). Esta faixa de horário foi utilizada, pois, em geral, representa horários onde a maior parte da população se encontra em suas residências, uma vez que este intervalo está fora do horário de jornada de trabalho habitual.

Após a execução da agregação pelo algoritmo DBSCAN, foi calculado o centro de massa (centroide) do maior *cluster* encontrado pelo algoritmo, sendo este considerado o local de residência. Para fins de processamento, este centroide é utilizado pelo método como ponto ao qual serão extraídos os indicadores sociais do LSOA que o contém, permitindo assim, o cálculo das correlações estatísticas entre estes indicadores e as variáveis de mobilidade (padrões de mobilidade) do indivíduo.

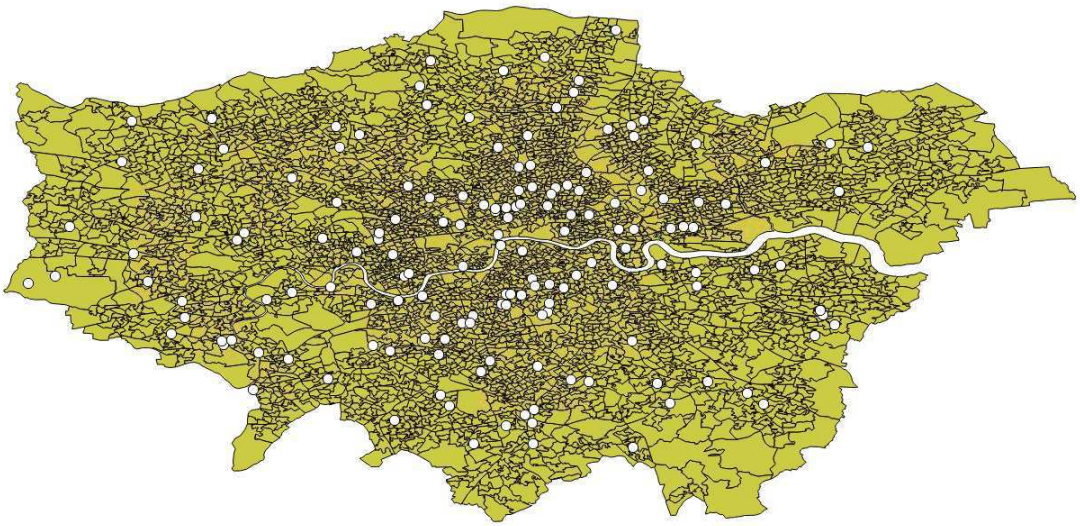
As Figuras de 22 a 24 mostram os locais de residências detectados pelo método para cada uma das três categorias citadas na Seção 5.2.1, onde foram observadas, para a Categoria 1, um total de 100.666 mensagens geradas a partir de residências; para a Categoria 2, foram detectadas 35.658 mensagens; e, para a Categoria 3, 3.506 mensagens.

Figura 22 – Residências detectadas para usuários com pelo menos 1000 tweets (Categoria 1)



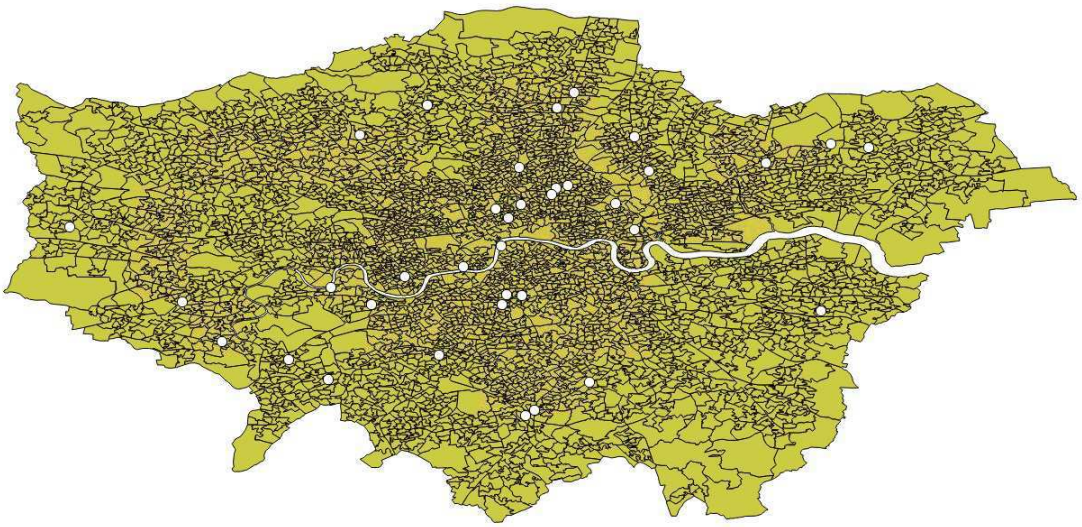
Fonte: Produzido pelo autor

Figura 23 – Residências detectadas para usuários com pelo menos 2500 tweets (Categoria 2)



Fonte: Produzido pelo autor

Figura 24 – Residências detectadas para usuários com pelo menos 5000 tweets (Categoria 3)



Fonte: Produzido pelo autor

Como é possível observar, as residências detectadas estão bem distribuídas sobre a área da cidade de Londres, representando uma condição importante à análise proposta, pois permite contemplar usuários de diferentes regiões da cidade, não se restringindo a localidades específicas, com características próprias, por exemplo.

5.3.1 Resultados do experimento

Como resultados extraídos das matrizes de correlação geradas pelo método aqui proposto, foram encontrados resultados relevantes apenas para a Categoria 3, descrita na Seção 5.2.1. As demais categorias (Categoria 1 e Categoria 2) não apresentaram correlações significativas a este trabalho.

Para os 36 usuários pertencentes à Categoria 3, totalizando 286.827 mensagens postadas, foi observado os seguintes valores para cada nível de filtragem adotado:

a) mensagens postadas apenas em feriados (*bank holidays*) que ocorreram na cidade de Londres durante o período de coleta das mensagens: 34 usuários e 9.369 mensagens postadas;

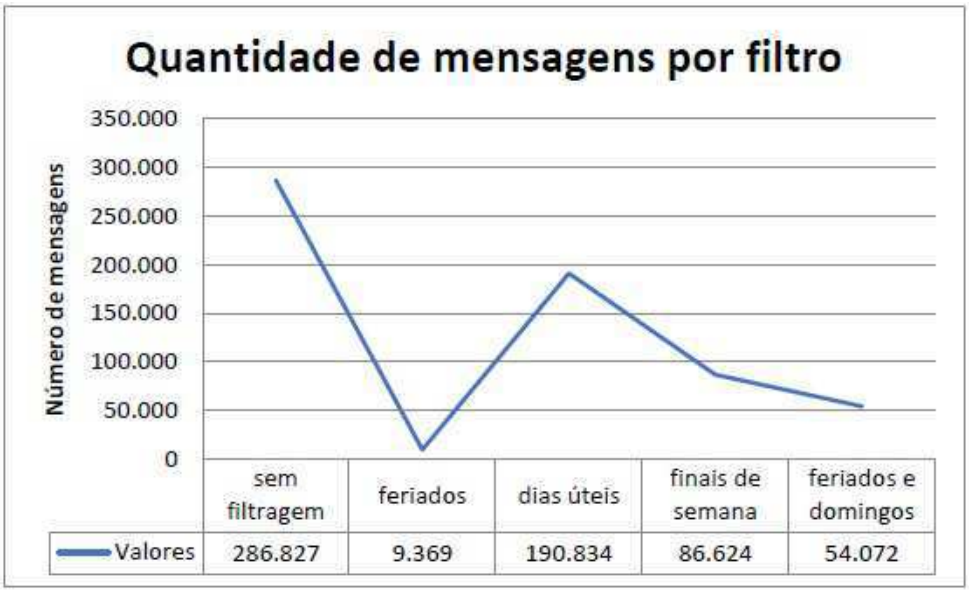
b) mensagens postadas apenas em dias úteis: 36 usuários e 190.834 mensagens encontradas;

c) mensagens postadas apenas durante os finais de semana (sábado e domingo): 36 usuários e 86.624 mensagens observadas.

d) mensagens postadas durante feriados e aos domingos: 36 usuários e 54.072 mensagens observadas;

Para os usuários desta categoria, a Figura 25 mostra a evolução no número de mensagens postadas por cada filtro implementado no método.

Figura 25 – Gráfico do número de mensagens postadas para cada um dos filtros utilizados



Fonte: Produzido pelo autor

As próximas subseções apresentam os resultados obtidos após a execução do método, apresentando tabelas que exibem as correlações entre uma variável de mobilidade e as variáveis sociais correlacionadas a esta. As tabelas possuem seus valores exibidos em forma de tuplas (τ, *p-value*), onde τ representa o coeficiente de correlação de Kendall e o *p-value* representa a significância estatística obtida pelo teste de correlação executado.

5.3.1.1 Resultados do Experimento 1 para a Q2

Para a execução dos experimentos dentro do escopo da Q2, as tabelas desta seção apresentam os resultados obtidos para cada variável de mobilidade e suas variáveis sociais correlacionadas pelo método.

Para a análise da variável de mobilidade "Raio de Giro"(Tabela 8), foram detectadas correlações com treze variáveis sociais. Para a segunda coluna da tabela, onde não há nenhuma filtragem temporal, pode-se observar a correlação positiva relacionada à variável social “Taxa de empregabilidade” ($\tau = 0{,}27$), sugerindo que indivíduos que possuem um Raio de Giro maior tendem a morar em regiões com uma taxa de empregabilidade maior. Este comportamento pode ser justificado pelo fato de que pessoas que possuem uma rotina de trabalho diária tendem a realizar deslocamentos maiores e regulares.

Ainda nesta coluna, correlações negativas foram encontradas para a variável "Pessoas sem qualificações profissionais" (τ = -0,26), sugerindo que, quanto maior o Raio de Giro de um indivíduo, o índice de pessoas sem qualificações profissionais na região de sua residência tende a ser menor. O mesmo padrão foi encontrado para as variáveis "Total de pessoas economicamente inativas" (τ = -0,30) e "Muçulmanos" (τ = -0,31).

Foi observado um comportamento similar entre os demais níveis de filtragem. Por exemplo, as correlações relacionadas à variável "Total de pessoas economicamente inativas" sugerem que pessoas que moram em regiões onde há um valor maior para este indicador, possuem um Raio de Giro menor ao se analisar isoladamente mensagens postadas nos feriados e domingos (τ = -0,26) e nos dias úteis (τ = -0,34).

Tabela 8 – Correlações encontradas para usuários da Categoria 3 (Raio de Giro)

Variável Social	Sem filtragem	Feriados (Nível 1)	Feriados + Domingos (Nível 2)	Dias úteis (Nível 3)	Finais de semana (Nível 4)
Total de pessoas economicamente inativas	(-0.30, 0.01)	-	(-0.26, 0.02)	(-0.34, 0.003)	-
Taxa de empregabilidade	(0.27, 0.01)	-	-	(0.27, 0.01)	-
Pessoal sem qualificações profissionais	(-0.26, 0.02)	-	(-0.37, 0.001)	-	(-0.38,0.0008)
Muçulmanos	(-0.31, 0.007)	-	(-0.30, 0.01)	(-0.29, 0.01)	(-0.30, 0.009)
Faixa etária de 0 a 15 anos	-	-	(-0.28, 0.01)	-	(-0.33, 0.004)
Hindu	-	(0.31, 0.009)	-	-	-
Outras religiões	-	(0.30, 0.01)	-	-	-
Sales	(0.27, 0.02)	-	-	-	-
Pessoas economicamente ativas desempregadas	-	(-0.26, 0.03)	(-0.32, 0.005)	-	(-0.28, 0.01)
Taxa de desemprego	-	(-0.25, 0.03)	(-0.29, 0.01)	-	(-0.25, 0.02)
Múltiplos grupos étnicos	-	-	(-0.25, 0.03)	-	-
Negros, Africanos, Caribenhos e Negros britânicos	-	-	(-0.29, 0.01)	-	(-0.28, 0.01)
Atividades diárias um pouco limitadas por condições físicas	-	-	(-0.28, 0.01)	-	(-0.25, 0.02)

Fonte: Produzido pelo autor

Para a variável de mobilidade "Total de Distância Percorrida"(Tabela 9), na segunda coluna, onde novamente não foram feitas quaisquer filtragens temporais, é possível destacar as correlações encontradas para as variáveis sociais “Pessoas economicamente ativas desempregadas” (τ = -0,33), “Pessoas sem qualificações profissionais” (τ = -0,33) e “Estudantes em tempo integral economicamente ativos” (τ = -0,25). Para estas variáveis sociais, quanto maior o valor da variável de mobilidade Total de Distância Percorrida para um indivíduo, menor será o valor destes indicadores para a sua região de residência, dado a negatividade da correlação. Este resultado confirma a tendência de que fatores relacionados à empregabilidade (ou a falta dela) parecem exercer algum nível de influência nas distâncias percorridas por um indivíduo.

Tabela 9 – Correlações encontradas para usuários da Categoria 3 (Total de Distância Percorrida)

Variável Social	Sem filtragem	Feriados (Nível 1)	Feriados + Domingos (Nível 2)	Dias úteis (Nível 3)	Finais de semana (Nível 4)
Pessoas economicamente ativas desempregadas	(-0.33, 0.003)	-	(-0.32, 0.004)	(-0.29, 0.01)	(-0.33, 0.004)
Estudantes em tempo integrar economicamente ativos	(-0.25, 0.02)	-	-	-	-
Pessoas sem qualificações profissionais	(-0.33, 0.004)	(-0.28, 0.01)	(-0.40,0.0005)	(-0.28, 0.01)	(-0.39,0.0006)
Muçulmanos	(-0.29, 0.01)	-	-	(-0.27, 0.01)	(-0.25, 0.03)
Atividades diárias um pouco limitadas por condições físicas	(-0.26, 0.02)	(-0.33, 0.006)	(-0.40,0.0005)	-	(-0.36, 0.001)
Atividades diárias muito limitadas por condições físicas	-	(-0.25, 0.03)	(-0.27, 0.01)	-	-
Outras religiões	-	(0.25, 0.03)	-	-	-
Faixa etária de 0 a 15 anos	(-0.29, 0.01)	-	(-0.38,0.0008)	-	(-0.43,0.0002)
Faixa etária de 45 a 64 anos	-	-	(-0.27, 0.01)	-	(-0.28, 0.01)
Total de pessoas economicamente inativas	(-0.25, 0.03)	-	-	(-0.26, 0.02)	-
Atividades diárias não limitadas por condições físicas	-	-	-	-	(-0.25, 0.03)
Múltiplos grupos étnicos	(-0.29, 0.01)	-	(-0.25, 0.03)	-	-
Negros, Africanos, Caribenhos e Negros britânicos	(-0.28, 0.01)	-	(-0.26, 0.02)	(-0.26, 0.02)	(-0.27, 0.01)
Taxa de desemprego	(-0.27, 0.01)	-	(-0.25, 0.02)	(-0.26, 0.02)	(-0.26, 0.02)
Pessoas nascidas no Reino Unido	-	-	(-0.31, 0.006)	-	(-0.28, 0.01)
Casais com filhos dependentes	-	-	-	-	(-0.27, 0.01)

Fonte: Produzido pelo autor

Para os demais níveis, têm-se como exemplo novamente a variável "Pessoas sem qualificações profissionais", onde esta também apresenta uma relação inversamente proporcional à variável "Total de Distância Percorrida" tanto para feriados e domingos (τ = -0,27) quanto para dias úteis (τ = -0,25) e finais de semana (τ = -0,26).

A variável de mobilidade "Número de Deslocamentos" (Tabela 10) apresentou poucos resultados, revelando correlação apenas com a variável social "Famílias com quatro ou mais carros ou vans". As correlações obtidas sugerem que pessoas que possuem um valor maior para esta variável de mobilidade tenderiam a morar em regiões que possuem uma quantidade menor de indivíduos possuindo mais de quatro veículos. Os resultados para esta variável sugerem que o número de deslocamentos não está associado, necessariamente, à quantidade de automóveis presentes em uma residência, pelo menos neste estudo.

Tabela 10 – Correlações encontradas para usuários da Categoria 3 (Número de Deslocamentos)

Variável Social	Sem filtragem	Feriados (Nível 1)	Feriados + Domingos (Nível 2)	Dias úteis (Nível 3)	Finais de semana (Nível 4)
Famílias com quatro ou mais carros ou vans	(-0.28, 0.02)	-	-	(-0.26, 0.02)	(-0.27, 0.02)

Fonte: Produzido pelo autor

Para a variável de mobilidade "Média de Deslocamentos Por Dia" (Tabela 11), são apresentados resultados semelhantes aos encontrados na Tabela 10.

Tabela 11 – Correlações encontradas para usuários da Categoria 3 (Média de Deslocamentos Por Dia)

Variável Social	Sem filtragem	Feriados (Nível 1)	Feriados + Domingos (Nível 2)	Dias úteis (Nível 3)	Finais de semana (Nível 4)
Famílias com quatro ou mais carros ou vans	(-0.34, 0.005)	-	-	-0.33, 0.006)	(-0.25, 0.03)

Fonte: Produzido pelo autor

A variável de mobilidade "Média de Distância Entre Deslocamentos" (Tabela 12) apresenta um comportamento diferente do apresentado pelas duas últimas variáveis de mobilidade, em especial para a variável "Famílias com quatro ou mais carros ou vans". Nestes resultados, foram observadas correlações positivas, sugerindo que famílias que possuem mais carros em suas garagens parecem se deslocar por maiores distâncias entre deslocamentos, tanto nos feriados e domingos (τ = 0,26), como nos dias úteis (τ = 0,35) e finais de semana (τ = 0,25).

Tabela 12 – Correlações encontradas para usuários da Categoria 3 (Média de Distância Entre Deslocamentos)

Variável Social	Sem filtragem	Feriados (Nível 1)	Feriados + Domingos (Nível 2)	Dias úteis (Nível 3)	Finais de semana (Nível 4)
Famílias com quatro ou mais carros ou vans	(0.33, 0.006)	-	(0.26, 0.03)	(0.35, 0.004)	(0.25, 0.03)
Famílias com três ou mais carros ou vans	-	-	-	-	-
Muçulmanos	(-0.33, 0.004)	-	(-0.37, 0.001)	(-0.28, 0.01)	(-0.32, 0.006)
Em idade de trabalho	-	-	(-0.26, 0.02)	-	-
Total de pessoas economicamente inativas	(-0.25, 0.02)	-	(-0.27, 0.01)	(-0.25, 0.02)	(-0.26, 0.02)
Pessoas economicamente ativas desempregadas	(-0.29, 0.01)	-	(-0.38, 0.001)	(-0.26, 0.02)	(-0.30, 0.01)
Taxa de desemprego	(-0.27, 0.01)	-	(-0.32, 0.005)	-	(-0.28, 0.01)
Pessoas sem qualificações profissionais	-	-	(-0.29, 0.01)	-	(-0.27, 0.02)
Faixa etária de 16 a 29 anos	-	-	-	-	-
Faixa etária de 0 a 15 anos	(-0.25, 0.02)	-	(-0.31, 0.006)	-	(-0.33, 0.004)
Atividades diárias não limitadas por condições físicas	-	-	(-0.26, 0.02)	-	(-0.27, 0.01)
Boa saúde	-	-	(-0.27, 0.02)	-	(-0.27, 0.02)
Negros, Africanos, Caribenhos e Negros britânicos	(-0.25, 0.03)	-	(-0.32, 0.005)	(-0.25, 0.02)	(-0.26, 0.02)
Múltiplos grupos étnicos	-	-	(-0.25, 0.03)	-	-
Atividades diárias um pouco limitadas por condições físicas	-	-	(-0.27, 0.01)	-	(-0.27, 0.01)

Fonte: Produzido pelo autor

É importante destacar também as correlações relacionadas às variáveis “Total de pessoas economicamente inativas”, “Taxa de desemprego” e “Pessoas sem qualificações profissionais”. Os resultados obtidos na Tabela 12 mostram um relacionamento inverso

entre estes indicadores sociais e a variável de mobilidade em questão, o que sugere que indivíduos que possuem uma média de distância entre deslocamentos maior parecem habitar em regiões onde estes quatro indicadores sociais são menores. Estas correlações podem indicar, por exemplo, que pessoas economicamente inativas, desempregadas ou sem qualificações se deslocam por distâncias menores em praticamente todos os níveis de filtragens, possivelmente por possuírem uma renda menor. Comportamento análogo pode ser observado pelos indicadores "Negros/Africanos/Caribenhos/Negros britânicos" e "Múltiplos grupos étnicos".

A variável de mobilidade "Média de Preços de POI Visitados" (Tabela 13) apresenta correlação com oito indicadores sociais, dentre os quais, têm-se como destaque novamente a incidência de indicadores sociais relacionados a questões de empregabilidade, como o caso do indicador "Total de pessoas economicamente inativas". Para esse indicador, foi encontrado correlação negativa com mensagens postadas em dias úteis (τ = -0,28), sugerindo que pessoas que visitam POI com preços mais elevados, tendem a morar em regiões onde o indicador social "Total de pessoas economicamente inativas" é menor, considerando-se apenas mensagens postadas nos dias úteis.

Tabela 13 – Correlações encontradas para usuários da Categoria 3 (Média de Preços de POI Visitados)

Variável Social	Sem filtragem	Feriados (Nível 1)	Feriados + Domingos (Nível 2)	Dias úteis (Nível 3)	Finais de semana (Nível 4)
Faixa etária de 0 a 15 anos	-	(-0.31, 0.02)	-	-	-
Faixa etária de 45 a 64 anos	(-0.53, 0.0001)	-	-	(-0.47, 0.0007)	-
Casais com filhos dependentes	(-0.34, 0.01)	(-0.29, 0.04)	-	-	-
Pessoas nascidas no Reino Unido	(-0.30, 0.02)	(-0.35, 0.01)	-	-	-
Total de pessoas economicamente inativas	-	-	-	(-0.28, 0.04)	-
Atividades diárias um pouco limitadas por condições físicas	(-0.29, 0.03)	-	-	-	-
Boa saúde	(-0.28, 0.04)	-	-	-	-
Adeptos do Siquismo	-	-	-	-	(0.31, 0.02)

Fonte: Produzido pelo autor

No experimento, também foram observadas correlações entre a variável de mo-

bilidade e o indicador social "Casais com filhos dependentes". Para esse indicador, as correlações encontradas foram sempre negativas (mensagens sem filtragem temporal e mensagens postadas nos feriados), indicando uma tendência de que usuários que visitam POI com preços maiores tendem a residir em regiões da cidade onde o indicador social "Casais com filhos dependentes" é menor.

5.3.1.2 Discussão dos resultados para o Experimento 1

Para o Experimento 1, que visa responder à questão Q2, a qual indaga sobre a existência de correlações estatísticas entre padrões de mobilidade e os indicadores sociais presentes na região de residência dos usuários, foram verificadas correlações em diversos ensaios executados. Para os casos onde não foram executadas nenhum nível de filtragem temporal, foram encontradas correlações estatísticas relacionadas principalmente às condições de trabalho. Pode-se destacar as variáveis "Taxa de empregabilidade" (Tabela 8) e "Pessoas sem qualificações profissionais" (Tabela 9), as quais se correlacionaram com diversas variáveis de mobilidade e apresentaram comportamentos concordantes, sugerindo que pessoas que se deslocam sob maiores distâncias tendem a residir em locais onde a taxa de empregabilidade é maior e a taxa de pessoas sem qualificações profissionais é menor.

Comportamento similar também pode ser observado ao se analisar a variável "Pessoas economicamente ativas desempregadas", que tende a diminuir quando a variável "Total de Distância Percorrida" aumenta (Tabela 9).

Padrões relacionados à renda também podem ser visíveis ao se analisar a variável de mobilidade "Media de preços de POI visitados" (Tabela 13), a qual se correlaciona negativamente com os indicadores sociais "Total de pessoas economicamente inativas" e "Casais com filhos dependentes". Esse resultado sugere que usuários que visitam POI mais caros tendem a residir em regiões onde esses dois indicadores sociais são menores. O resultado apresentado pode estar relacionado com uma possível restrição de renda tanto de pessoas economicamente inativas quanto para casais com filhos dependentes.

O comportamento apresentado por variáveis de cunho étnico/religioso se repete ao longo dos experimentos, onde indicadores que expressam populações estrangeiras parecem demonstrar que estes grupos executam deslocamentos menores. Esse resultado pode demonstrar uma possível segregação destas populações, onde pessoas de um determinado grupo étnico se deslocariam mais por determinados bairros ou regiões da cidade. Esse resultado concorda com o trabalho de Luo et al. (2016), onde os autores também identificaram um comportamento semelhante ao considerar os padrões de mobilidade de pessoas de diferentes etnias na cidade de Chicago. A Tabela 14 exibe os principais resultados obtidos para o Experimento 1.

Tabela 14 – Principais resultados encontrados para o Experimento 1

Variável social	Variável de mobilidade	Sem filtragem	Feriados (Nível 1)	Feriados + Domingos (Nível 2)	Dias úteis (Nível 3)	Finais de semana (Nível 4)
Taxa de empregabilidade	Raio de Giro	(0.27, 0.01)	-	-	(0.27, 0.01)	-
Pessoas sem qualificações profissionais	Raio de Giro	(-0.26, 0.02)	-	(-0.37, 0.001)	-	(-0.38,0.0008)
Total de pessoas economicamente inativas	Raio de Giro	(-0.30, 0.01)	-	(-0.26, 0.02)	(-0.34, 0.003)	-
Muçulmanos	Raio de Giro	(-0.31, 0.007)	-	(-0.30, 0.01)	(-0.29, 0.01)	(-0.30, 0.009)
Pessoas economicamente ativas desempregadas	Total de Distância Percorrida	(-0.33, 0.003)	-	(-0.32, 0.004)	(-0.29, 0.01)	(-0.33, 0.004)
Pessoas sem qualificações profissionais	Total de Distância Percorrida	(-0.33, 0.004)	(-0.28, 0.01)	(-0.40,0.0005)	(-0.28, 0.01)	(-0.39,0.0006)
Taxa de desemprego	Média de Distância Entre Deslocamentos	(-0.27, 0.01)	-	(-0.32, 0.005)	-	(-0.28, 0.01)
Negros, Africanos, Caribenhos e Negros britânicos	Média de Distância Entre Deslocamentos	(-0.25, 0.03)	-	(-0.32, 0.005)	(-0.25, 0.02)	(-0.26, 0.02)
Total de pessoas economicamente inativas	Média de Preços de POI Visitados	-	-	-	(-0.28, 0.04)	-

Fonte: Produzido pelo autor

5.4 Experimento 2: análise de correlação entre padrões de mobilidade e regiões visitadas

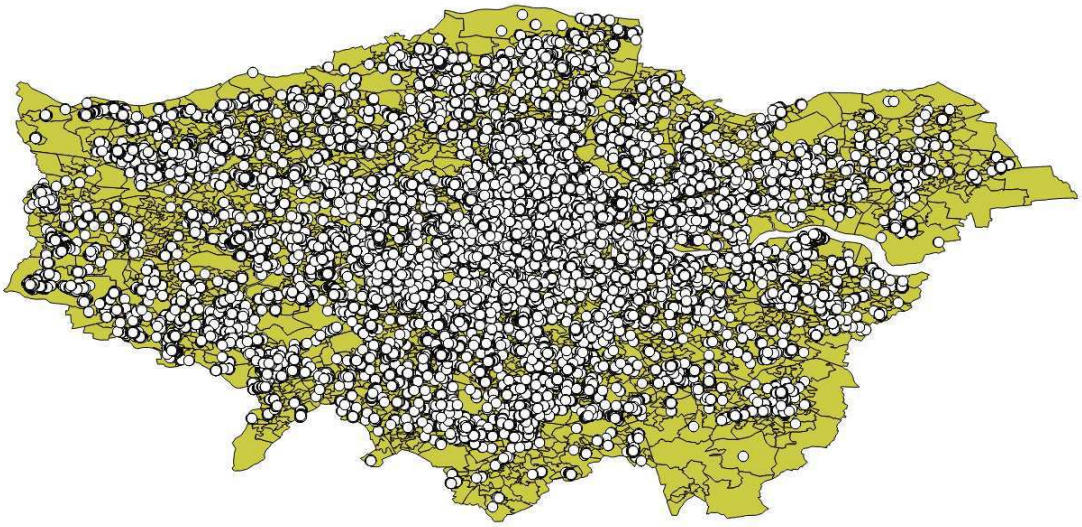
O segundo experimento conduzido neste trabalho visa verificar a existência de correlações entre os padrões de mobilidade de um indivíduo e os indicadores sociais presentes nas regiões dos AC do indivíduo. Este experimento é relevante pois tenta verificar se determinados padrões de mobilidade estão associados a visitas em regiões com determinadas características sociais. Por exemplo, um usuário que se desloca por distâncias maiores durante o dia tende a visitar regiões onde a taxa de desemprego é maior?

Para analisar questões deste tipo, foi utilizando o algoritmo DBSCAN para agregar as mensagens postadas por um usuário em AC, onde foram calculadas as medianas para

cada indicador social associado às regiões onde os *clusters* foram formados. Neste caso em particular, a mediana foi utilizada devido a alta incidência de *outliers* observados nos indicadores sociais.

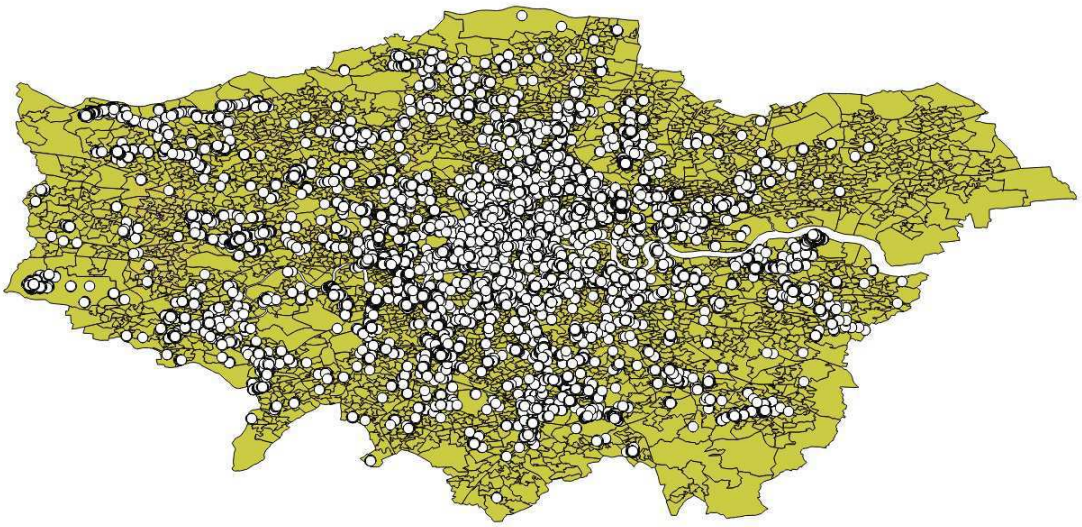
As Figuras 26, 27 e 28 mostram os AC detectados pelo método para cada uma das três categorias analisadas neste trabalho.

Figura 26 – AC para usuários com pelo menos 1.000 mensagens postadas



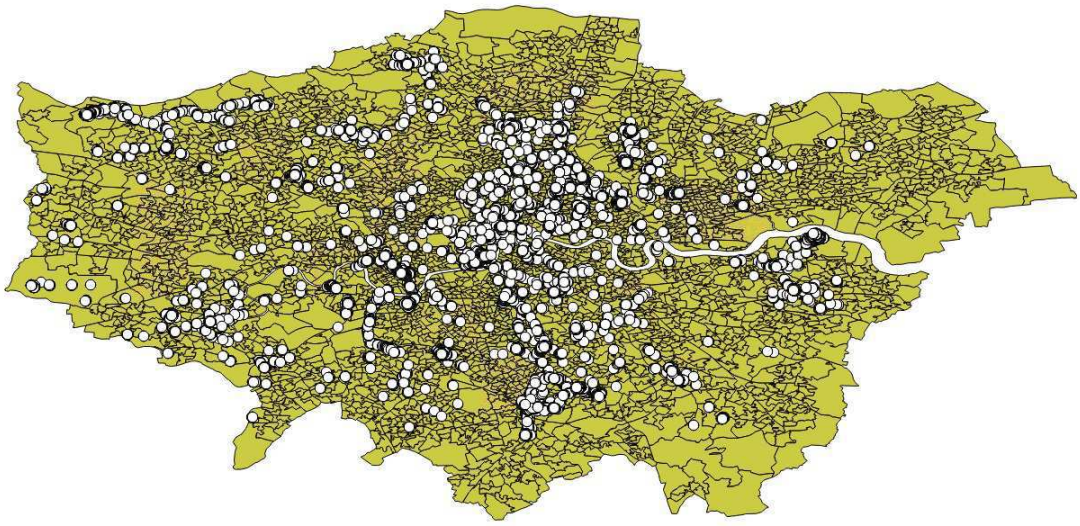
Fonte: Produzido pelo autor

Figura 27 – AC para usuários com pelo menos 2.500 mensagens postadas



Fonte: Produzido pelo autor

Figura 28 – AC para usuários com pelo menos 5.000 mensagens postadas



Fonte: Produzido pelo autor

5.4.1 Resultados do experimento

Para este experimento proposto, não foram encontradas correlações estatísticas significativas para as Categorias 1 e 2. Assim como no primeiro experimento, apenas a Categoria 3 apresentou correlações significativas.

As próximas subseções apresentam os resultados obtidos após a execução do método, apresentando tabelas que exibem as correlações entre uma variável de mobilidade e as variáveis sociais correlacionadas a esta. As tabelas, assim como no primeiro experimento demonstrado neste trabalho, possuem seus valores exibidos em forma de tuplas (τ, *p-value*), onde τ representa o coeficiente de correlação de Kendall e o *p-value* representa a significância estatística obtida pelo teste de correlação executado.

5.4.1.1 Resultados obtidos para o Experimento 2 para a Q3

Como resposta para a pergunta de pesquisa Q3, que visa verificar possíveis correlações entre os padrões de mobilidade e os indicadores sociais das regiões frequentemente visitadas, temos nesta seção os resultados obtidos para esse experimento.

Analisando a variável de mobilidade “Raio de Giro” (Tabela 15), podemos destacar na segunda coluna (dados sem filtragem temporal) a variável social “Pessoas sem qualificações profissionais” (τ = -0,31), onde este resultado sugere que quanto maior o Raio de Giro de um indivíduo, ele tenderá a se deslocar com mais frequência por regiões com um menor valor para esta variável social, inclusive nos dias úteis (τ = -0,29). Esse resultado

concorda com os resultados obtidos para o Experimento 1, onde o Raio de Giro também está correlacionado com esta variável social.

Tabela 15 – Correlações encontradas para usuários da Categoria 3 - Q3 (Raio de Giro)

Variável Social	Sem filtragem	Feriados (Nível 1)	Feriados + Domingos (Nível 2)	Dias úteis(Nível 3)	Finais de semana (Nível 4)
Pessoas sem qualificações profissionais	(-0.31, 0.007)	-	-	(-0.29, 0.01)	-
Pessoas brancas	-	(-0.25, 0.03)	-	-	-
Pessoas economicamente ativas desempregadas	-	-	(-0.27, 0.01)	-	-
Pessoas economicamente ativas empregadas	-	(-0.26, 0.02)	-	-	-
Sem religião	-	(-0.33, 0.005)	-	-	-
Faixa etária de 16 a 29 anos	-	(-0.25, 0.03)	-	-	-
Faixa etária de 30 a 44 anos	-	(-0.30, 0.01)	-	-	-
Em idade de trabalho	-	(-0.38, 0.001)	(-0.28, 0.01)	-	-
Pessoas nascidas no Reino Unido	-	(-0.36, 0.002)	-	-	-
Total de pessoas economicamente ativas	-	(-0.30, 0.01)	-	-	-
Hinduístas	-	(0.29, 0.01)	-	-	-
Outras religiões	-	(0.25, 0.04)	-	-	-

Fonte: Produzido pelo autor

Para a variável de mobilidade “Total de Distância Percorrida” (Tabela 16), na segunda coluna (sem filtragem temporal), é possível verificar a correlação desta variável com diversos indicadores sociais, porém é verificada novamente a incidência de variáveis relacionadas a condições de trabalho bem como a variáveis relacionadas a pessoas imigrantes. Como exemplo, tem-se o indicador social “Taxa de empregabilidade”, que apresenta uma correlação positiva (τ = 0,26) com a variável de mobilidade analisada, demonstrando que valores maiores para esta variável de mobilidade estão correlacionados com AC formados em regiões onde a taxa de empregabilidade é maior. Concordando com essa variável, porém em sentido oposto, tem-se o resultado, ainda na segunda coluna, da variável “Pessoas economicamente ativas desempregadas” (τ = -0,25), sugerindo que pessoas que se deslocam por distâncias maiores tendem a visitar regiões onde este indicador social é mais baixo, inclusive nos feriados (τ = -0,27) e finais de semana (τ = -0,27).

Tabela 16 – Correlações encontradas para usuários da Categoria 3 - Q3 (Total de Distância Percorrida)

Variável Social	Sem filtragem	Feriados (Nível 1)	Feriados + Domingos (Nível 2)	Dias úteis (Nível 3)	Finais de semana (Nível 4)
Pessoas economicamente ativas desempregadas	(-0.25, 0.02)	(-0.27, 0.02)	(-0.27, 0.01)	-	(-0.27, 0.01)
Taxa de empregabilidade	(0.26, 0.02)	-	-	(0.27, 0.01)	-
Pessoas sem qualificações profissionais	(-0.28, 0.01)	-	-	(-0.28, 0.01)	-
Atividades diárias muito limitadas por condições físicas	(-0.30, 0.009)	-	-	(-0.30, 0.008)	-
Faixa etária de 16 a 29 anos	-	(-0.33, 0.005)	-	(-0.26, 0.02)	-
Faixa etária de 30 a 44 anos	-	(-0.35, 0.003)	-	-	-
Em idade de trabalho	-	(-0.45, 0.0001)	-	-	-
Total de famílias	-	(-0.25, 0.03)	-	-	-
Pessoas nascidas no Reino Unido	-	(-0.29, 0.01)	-	-	-
Total de pessoas economicamente ativas	-	(-0.33, 0.005)	-	-	-
Pessoas economicamente ativas empregadas	-	(-0.27, 0.02)	-	-	-
Estudantes em tempo integrar economicamente ativos	-	(-0.27, 0.02)	-	-	-
Atividades diárias não limitadas por condições físicas	-	(-0.29, 0.01)	-	-	-
Boa saúde	-	(-0.29,0.01)	-	-	-
Sem religião	-	(-0.29, 0.01)	-	-	-
Muçulmanos	-	-	-	(-0.25, 0.03)	-
Famílias com três ou mais carros ou vans	-	-	-	(0.25, 0.03)	-

Fonte: Produzido pelo autor

A variável de mobilidade “Número de Deslocamentos” (Tabela 17) não apresentou

resultados na segunda coluna (dados sem filtragem temporal), apresentando resultados apenas para o Nível 1. Para esta variável de mobilidade, destacamos as variáveis "Em idade de trabalho" (τ = -0,32) e "Total de pessoas economicamente ativas" (τ = -0,26), onde quanto maior o número de deslocamentos para um indivíduo (variável de mobilidade), maiores serão as chances de este indivíduo visitar regiões com valores mais baixos para estes dois indicadores sociais nos feriados, dado as correlações negativas encontradas.

Tabela 17 – Correlações encontradas para usuários da Categoria 3 - Q3 (Número de Deslocamentos)

Variável Social	Sem filtragem	Feriados (Nível 1)	Feriados + Domingos (Nível 2)	Dias úteis (Nível 3)	Finais de semana (Nível 4)
Faixa etária de 30 a 44 anos	-	(-0.26, 0.02)	-	-	-
Em idade de trabalho	-	(-0.32, 0.007)	-	-	-
Total de pessoas economicamente ativas	-	(-0.26, 0.02)	-	-	-

Fonte: Produzido pelo autor

A variável de mobilidade "Média de Deslocamentos Por Dia" (Tabela 18), apresentou correlação apenas com o indicador social "Média de preços de imóveis" (τ = -0,27), o que sugere uma relação inversamente proporcional para as duas variáveis, onde, quanto maior a variável de mobilidade, menor serão os preços dos imóveis das regiões visitadas durante os feriados. Resultado semelhante foi encontrado para a variável de mobilidade "Média de Distância Entre Deslocamentos" (Tabela 19), estando esta variável correlacionada inversamente com o indicador social "Estudantes em tempo integral economicamente ativos".

Tabela 18 – Correlações encontradas para usuários da Categoria 3 - Q3 (Média de Deslocamentos Por Dia)

Variável Social	Sem filtragem	Feriados (Nível 1)	Feriados + Domingos (Nível 2)	Dias úteis (Nível 3)	Finais de semana (Nível 4)
Média de preços de imóveis	-	(-0.27, 0.02)	-	-	-

Fonte: Produzido pelo autor

Tabela 19 – Correlações encontradas para usuários da Categoria 3 - Q3 (Média de Distância Entre Deslocamentos)

Variável Social	Sem filtragem	Feriados (Nível 1)	Feriados + Domingos (Nível 2)	Dias úteis (Nível 3)	Finais de semana (Nível 4)
Estudantes em tempo integrar economicamente ativos	(-0.26, 0.02)	-	-	-	-

Fonte: Produzido pelo autor

Para o indicador de mobilidade "Média de Preços de POI Visitados" (Tabela 20), dentre as correlações apresentadas, pode-se destacar a variável "Total de pessoas economicamente inativas". Esse indicador social apresentou correlação negativa com o indicador de mobilidade supracitado (τ = -0,30). Esse resultado pode sugerir que usuários que visitam POI com preços maiores tendem a visitar regiões da cidade onde o indicador social "Total de pessoas economicamente inativas" é menor, ao menos durante os dias úteis.

Tabela 20 – Correlações encontradas para usuários da Categoria 3 - Q3 (Média de Preços de POI Visitados)

Variável Social	Sem filtragem	Feriados (Nível 1)	Feriados + Domingos (Nível 2)	Dias úteis (Nível 3)	Finais de semana (Nível 4)
Pessoas nascidas no Reino Unido	(-0.29, 0.03)	(-0.40, 0.004)	-	-	-
Cristãos	-	(-0.29, 0.04)	-	-	-
Sem religião	-	(-0.29, 0.04)	-	-	-
Total de pessoas economicamente inativas	-	-	-	(-0.30, 0.03)	-

Fonte: Produzido pelo autor

5.4.1.2 Discussão dos resultados para o Experimento 2

Para este experimento, assim como os experimentos relacionados à questão de pesquisa Q2, foram observadas correlações relacionadas a questões de empregabilidade. Para este cenário, pode ser citado o indicador "Pessoas sem qualificações profissionais" (Tabela 15) e (Tabela 16), demonstrando que pessoas que realizam deslocamentos por maiores distâncias tendem a fazê-los para regiões onde a taxa de pessoas sem qualificação é menor. Confirmando este raciocínio, o indicador social "Taxa de empregabilidade" se relaciona com a variável de mobilidade Total de Distância Percorrida analisada na Tabela 16, indicando que, quanto maior o valor para a variável de mobilidade, maiores foram as incidências de deslocamentos para regiões com uma taxa de empregabilidade maior. Em

linhas gerais, estes resultados demonstram que indivíduos que efetuam deslocamentos por distâncias maiores, tendem a fazê-los para regiões onde a taxa de empregabilidade é maior e a taxa de pessoas sem qualificação é menor.

No tocante à variável de mobilidade Média de Preços de POI visitados (Tabela 20), também foram observados resultados em concordância com os obtidos em Q2. Aqui, foi observado uma relação inversamente proporcional entre essa variável de mobilidade e o indicador social "Total de pessoas economicamente inativas", sugerindo que usuários que frequentam POI com preços maiores tenderam a postar mensagens de regiões da cidade onde o indicador social "Total de pessoas economicamente inativas" era menor. Esse resultado evidencia novamente a correlação entre padrões de mobilidade e indicadores sociais relacionados a condições de emprego. A Tabela 21 exibe os principais resultados obtidos para o Experimento 2.

Tabela 21 – Principais resultados encontrados para o Experimento 2.

Variável social	Variável de mobilidade	Sem filtragem	Feriados (Nível 1)	Feriados + Domingos (Nível 2)	Dias úteis(Nível 3)	Finais de semana (Nível 4)
Pessoas sem qualificações profissionais	Raio de Giro	(-0.31, 0.007)	-	-	(-0.29, 0.01)	-
Taxa de empregabilidade	Total de Distância Percorrida	(0.26, 0.02)	-	-	(0.27, 0.01)	-
Pessoas economicamente ativas desempregadas	Total de Distância Percorrida	(-0.25, 0.02)	(-0.27, 0.02)	(-0.27, 0.01)	-	(-0.27, 0.01)
Em idade de trabalho	Número de Deslocamentos	-	(-0.32, 0.007)	-	-	-
Total de pessoas economicamente inativas	Média de Preços de POI Visitados	-	-	-	(-0.30, 0.03)	-

Fonte: Produzido pelo autor

5.5 Limitações dos resultados

Como demonstrado nas seções anteriores, o método proposto permitiu a descoberta de diversas correlações estatísticas entre os padrões de mobilidade e indicadores sociais referentes à cidade de Londres.

Porém, é importante destacar que os resultados reportados neste trabalho possuem limitações que ainda precisam ser considerados em estudos futuros, principalmente no que tange à precisão e possíveis vieses encontrados nos dados processados.

5.5.1 Limitação dos indicadores de mobilidade

Quanto aos indicadores de mobilidade, a principal limitação destes consiste no fato de serem colhidos a partir de mensagens georreferenciadas da rede social Twitter. Essa rede social permite que usuários postem mensagens em lugares específicos, como restaurantes e lojas, bem como em momentos de deslocamentos, como em ônibus e metrôs. Essa característica permite que usuários possam postar mensagens de forma fragmentada e descontínua, o que pode acarretar padrões pouco precisos, mesmo ao considerar usuários com muitas mensagens postadas. Como exemplo, podem-se ter usuários que postam mensagens apenas de suas residências e nos locais de trabalho, gerando assim, padrões pouco representativos.

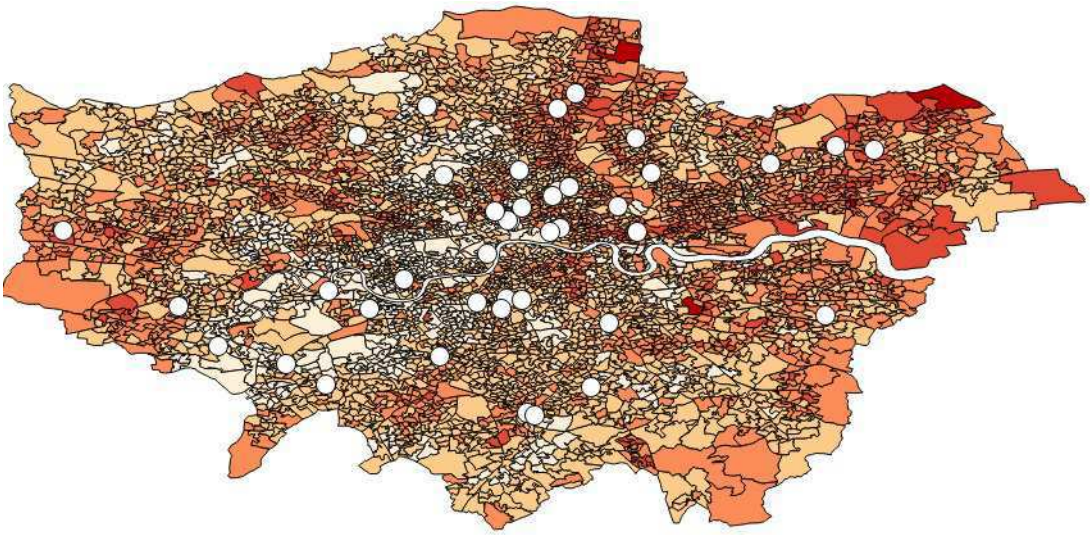
Ainda no tocante às dificuldades relacionadas aos dados do Twitter, é necessário considerar o quão representativo são os usuários utilizados neste estudo, posto que são apenas uma pequena amostra de um conjunto muito maior de moradores da cidade de Londres. Ainda como exemplo, pessoas muito pobres poderiam, em tese, utilizar menos a rede social, ou por possuírem aparelhos celulares mais modestos (sem GPS) ou mesmo por terem restrições ao acesso de redes 3G/4G.

5.5.2 Limitação dos indicadores sociais

Neste trabalho, os indicadores sociais representam valores vinculados a determinadas regiões geográficas da cidade de Londres. Nesse contexto, foi possível observar, por meio dos próprios dados da pesquisa, uma significante disparidade geográfica na distribuição dos indicadores sociais na área em estudo, o que pode levar à produção de vieses consideráveis nos resultados obtidos. Um exemplo de vieses desse tipo pode ser observados na Figura 29, que exibe, no mapa de Londres, graduações em cores para o indicador social "Pessoas sem qualificações profissionais", onde tons mais escuros representam um valor maior desse indicador social na região, e os pontos brancos representam as residências detectadas para usuários com pelo menos 5.000 mensagens postadas.

Na Figura 29, é possível observar que existem porções da cidade em que o indicador social possui valores maiores e menores, demonstrando desigualdades inerentes a grandes centros urbanos. Porém, essas desigualdades sociogeográficas observadas também podem contribuir para o enviesamento dos resultados apresentados neste trabalho. Como exemplo prático, é possível observar, na Figura 29, que a distribuição das residências não se encontra uniformemente distribuída sobre regiões com um maior e com menor valor para o indicador social representado na imagem, gerando assim, um possível viés quanto às correlações encontradas entre os padrões de mobilidade dos usuários e esse indicador social, o que pode ser estendido para todas as demais correlações encontradas neste trabalho.

Figura 29 – Graduação em cores para o indicador social "Pessoas sem qualificações profissionais"e as residências detectadas para usuários com pelo menos 5.000 mensagens postadas



Fonte: Produzido pelo autor

5.6 Discussão geral dos resultados

Por meio dos experimentos executados nesta pesquisa, foram encontradas 122 correlações para o Experimento 1 e 47 correlações para o Experimento 2. Dentre esses valores encontrados, foi observado um comportamento concordante entre correlações associadas, principalmente, a condições de trabalho e renda, como também com aspectos étnico-religiosos dos indivíduos.

Com os resultados apresentados neste trabalho, considerando tanto os resultados dos dois experimentos propostos, é possível verificar, por exemplo, que usuários que residem em regiões com uma maior taxa de empregabilidade tendem a se deslocar por maiores distâncias (Experimento 1), como também usuários que se deslocam por maiores distâncias tendem a visitar regiões onde existem menos pessoas economicamente ativas desempregadas (Experimento 2), demonstrando uma possível relação onde, pessoas com melhores condições de emprego e renda, tenderiam a visitar regiões com indicadores sociais similares aos delas.

Apesar dos resultados obtidos para as questões Q1, Q2 e Q3, que permitiriam rejeitar suas hipóteses nulas, verifica-se que, ao considerar os vieses discutidos na seção anterior, faz-se impraticável considerar que realmente existam correlações entre padrões de mobilidade e indicadores sociais relacionados aos locais de residência dos usuários, como também dos indicadores sociais relacionados aos seus AC, de modo a não rejeitar suas

hipóteses nulas, mesmo tendo sido encontradas correlações para esses dois casos. Para a real rejeição dessas hipóteses nulas, seria necessário um estudo mais amplo quanto ao real impacto que possíveis desigualdades sociais poderiam causar nos resultados. Um exemplo de possível viés, seria se, 80% das residências analisadas estivessem, na verdade, localizadas em regiões onde os moradores possuíssem um grande poder aquisitivo, tornando assim, os resultados substancialmente enviesados, comprometendo a análise estatística empregada.

Ao passo que os resultados desta pesquisa não permitem a rejeição das hipóteses nulas propostas, surgem questionamentos a serem abordados em pesquisas futuras. Novas hipóteses podem questionar o real impacto dos vieses citados sobre os resultados obtidos, bem como avaliar a real eficácia das novas métricas de mobilidade propostas. Também parece relevante empregar a técnica proposta nesta pesquisa com dados de trajetórias, e verificar se os resultados são similares aos encontrados utilizando-se mensagens de redes sociais, a exemplo do Twitter.

5.7 Considerações finais

Neste capítulo, foi apresentada a metodologia utilizada para a execução dos experimentos, visando a validação do método proposto por meio de três questões de pesquisa.

Os resultados obtidos, apesar de retornarem correlações estatisticamente significativas, foram considerados insuficientes para de fato rejeitar as hipóteses nulas propostas, dado um possível enviesamento dos dados sociais, cujos impactos não foram quantificados nesta pesquisa, impossibilitando assim a rejeição das hipóteses nulas presentes nesta pesquisa.

Por meio dos experimentos executados, foi observada a necessidade de se considerar um conjunto ainda maior de mensagens georreferenciadas para a extração de padrões de mobilidade. Isto se aplica, especialmente, a usuários com pelo menos 5.000 mensagens postadas, posto que esta classe de usuários foi a única a permitir a descoberta de correlações estatística entre os dados.

Quanto às quatro novas métricas de mobilidade propostas neste trabalho, foram observadas correlações altas entre duas delas ao longo dos experimentos, sendo estas o Número de Deslocamentos e o Número de Deslocamentos por dia. Essas duas métricas chegaram a apresentar correlações estatísticas de $\tau = 0.71$ entre si, o que podem sugerir uma possível redundância entre elas.

No próximo capítulo, serão apresentadas as considerações finais deste trabalho, bem como apontamentos para trabalhos futuros.

Parte VI

Conclusão

6 Conclusão

Mensagens postadas em redes sociais possuem a capacidade de agregar diferentes categorias de informações, ampliando as possibilidades de pesquisas científicas e análises sobre esses dados. Dentro desse contexto, aliado à tendência de modernização dos centros urbanos e visando uma melhoria na qualidade de vida de seus habitantes, faz-se imprescindível o aproveitamento dessa enorme massa de dados para a geração de conhecimento capaz de auxiliar a tomada de decisão por parte dos gestores e órgãos governamentais. O uso de dados de redes sociais com esse objetivo pode proporcionar grandes economias para governos, especialmente pelo baixo custo de obtenção destas informações, se comparado, por exemplo, com custos empregados em censos ou demais pesquisas em campo.

Esta pesquisa teve como objetivo desenvolver um método computacional capaz de extrair padrões de mobilidade a partir de mensagens de uma rede social e correlacionar estes padrões com indicadores sociais, permitindo observar, de forma objetiva, como estas duas classes de variáveis se relacionam.

O método apresentado permite uma análise automatizada dos dados, necessitando apenas de um conjunto de mensagens georreferenciadas de uma determinada região (e.g., uma cidade) e de uma tabela de banco de dados contendo indicadores sociais e polígonos referentes à área de estudo. Com estes dados, qualquer gestor pode, em pouco tempo, obter todos os coeficientes de correlação entre padrões de mobilidade e indicadores sociais, favorecendo uma visão mais ampla e analítica acerca do espaço urbano que o gestor administra. Os dados gerados pelo método aqui proposto também são pertinentes a sistemas de recomendação, provendo dados que podem aprimorar, de forma substancial, a recomendação de POI a indivíduos que residem em determinadas regiões da cidade, por exemplo.

Para o desenvolvimento do método aqui descrito, foram consideradas métricas para a análise de padrões de mobilidade já utilizadas na literatura, porém, também foram propostas novas métricas, permitindo um estudo mais amplo sobre suas possíveis relações com indicadores sociais.

O restante deste capítulo está organizado como segue: A Seção 6.1 descreve as principais contribuições geradas pela pesquisa apresentada nesta dissertação. A Seção 6.2 detalha os trabalhos futuros que podem ser realizados para ampliar a pesquisa.

6.1 Contribuições

O método aqui apresentado buscou atender a demandas ainda não supridas no tocante a padrões de mobilidade e suas relações com indicadores de cunho social. O Capítulo 3 demonstra esta necessidade, onde grande parte dos trabalhos não consideram aspectos sociais em seus estudos, e os que consideram, o fazem de forma bastante limitada. Além de propor uma abordagem analítica sobre padrões de mobilidade e indicadores sociais de forma mais ampla, este estudo buscou desenvolver um método automatizado para esta análise, facilitando, por exemplo, a tomada de decisões por parte de órgãos governamentais, bem como o enriquecimento de sistemas de recomendação.

Como principais contribuições desta pesquisa, citam-se:

a) Método computacional capaz de extrair padrões de mobilidade de mensagens de redes sociais e correlacionar esses padrões com indicadores sociais de uma região. Como resultado, foram identificadas diversas correlações, principalmente com indicadores associados a condições de trabalho e renda e grupos étnico-religiosos dos indivíduos;

b) Desenvolvimento de novas métricas e métodos para análise e estudo de padrões de mobilidade;

c) Integração de elementos como POI, AC e residências para a análise de padrões de mobilidade;

d) Implementação de diversas técnicas de filtragem de dados para a análise de mensagens de redes sociais.

6.2 Trabalhos futuros

Para trabalhos futuros, considerando as limitações encontradas neste estudo e visando ampliar a contribuição científica aqui apresentada, citam-se:

a) Executar o método com dados de outros centros urbanos, buscando eventuais semelhanças e diferenças entre os resultados obtidos;

b) Considerar utilizar mais usuários com pelo menos 5.000 mensagens postadas, dado que os resultados mais relevantes foram encontrados apenas neste grupo de usuários, reduzindo assim riscos relacionados a dados viesados;

c) Utilização de outras fontes de dados para a extração de POI. Neste trabalho, foram considerados apenas estabelecimentos reportados pela API do Foursquare;

d) Desenvolver uma interface gráfica que permita, de forma dinâmica, visualizar os resultados reportados pelo método. Atualmente, o método retorna apenas a matriz de correlação em formato *.xls*.

e) Considerar o estudo em diferentes granularidades geográficas. Neste trabalho, foi considerado apenas a divisão relativa a LSOA, podendo ser adotadas também divisões que considerem, por exemplo, os distritos de Londres;

f) Incorporar no método o uso do conteúdo das mensagens para a realização de análises sobre estas, correlacionando, por exemplo, possíveis tópicos e seus relacionamentos, tanto com padrões de mobilidade quanto com os indicadores sociais fornecidos.

Referências

ADEDOYIN-OLOWE, M. et al. A rule dynamics approach to event detection in twitter with its application to sports and politics. *Expert Systems with Applications*, v. 55, p. 351 – 360, 2016. ISSN 0957-4174. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0957417416300598>. Citado na página 29.

ALBUQUERQUE, J. P. de et al. A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management. *International Journal of Geographical Information Science*, v. 29, n. 4, p. 667–689, 2015. Disponível em: <http://dx.doi.org/10.1080/13658816.2014.996567>. Citado na página 30.

BAGROW, J. P.; LIN, Y.-R. Mesoscopic structure and social aspects of human mobility. *PloS one*, Public Library of Science, v. 7, n. 5, p. e37676, 2012. Citado na página 18.

BARBIER, G.; LIU, H. Data mining in social media. In: _____. *Social Network Data Analytics*. Boston, MA: Springer US, 2011. p. 327–352. ISBN 978-1-4419-8462-3. Disponível em: <http://dx.doi.org/10.1007/978-1-4419-8462-3_12>. Citado na página 29.

BARBOSA, L.; FENG, J. Robust sentiment detection on twitter from biased and noisy data. In: *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010. (COLING '10), p. 36–44. Disponível em: <http://dl.acm.org/citation.cfm?id=1944566.1944571>. Citado na página 29.

BATTY, M. et al. Smart cities of the future. *The European Physical Journal Special Topics*, v. 214, n. 1, p. 481–518, 2012. Citado na página 18.

BIRKIN, M. et al. An examination of personal mobility patterns in space and time using twitter. *International Journal of Agricultural and Environmental Information Systems (IJAEIS)*, IGI Global, v. 5, n. 3, p. 55–72, 2014. Citado 2 vezes nas páginas 38 e 52.

BLANFORD, J. I. et al. Geo-located tweets. enhancing mobility maps and capturing cross-border movement. *PloS one*, Public Library of Science, v. 10, n. 6, p. e0129202, 2015. Citado na página 38.

CHAINTREAU, A. et al. Impact of human mobility on opportunistic forwarding algorithms. *IEEE Transactions on Mobile Computing*, IEEE, v. 6, n. 6, p. 606–620, 2007. Citado na página 20.

CHEN, C.-C.; CHIANG, M.-F.; PENG, W.-C. Mining and clustering mobility evolution patterns from social media for urban informatics. *Knowledge and Information Systems*, Springer, v. 47, n. 2, p. 381–403, 2016. Citado na página 37.

CHENG, Z. et al. Exploring millions of footprints in location sharing services. *ICWSM*, v. 2011, p. 81–88, 2011. Citado 3 vezes nas páginas 40, 57 e 58.

CHOK, N. S. *Pearson's versus Spearman's and Kendall's correlation coefficients for continuous data*. Tese (Doutorado) — University of Pittsburgh, 2010. Citado 2 vezes nas páginas 32 e 61.

COHEN, J. Statistical power analysis for the behavioral sciences lawrence earlbaum associates. *Hillsdale, NJ*, p. 20–26, 1988. Citado na página 72.

CRANSHAW, J. et al. The livehoods project: Utilizing social media to understand the dynamics of a city. *International AAAI Conference on Weblogs and Social Media*, 2012. Citado na página 40.

DREDZE, M. et al. Twitter as a source of global mobility patterns for social good. *arXiv preprint arXiv:1606.06343*, 2016. Citado na página 39.

ESTER, M. et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. Menlo Park, California: AAAI Press, 1996. v. 96, p. 226–231. Citado na página 52.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. *AI magazine*, v. 17, n. 3, p. 37, 1996. Citado 4 vezes nas páginas 25, 26, 27 e 28.

FERRARI, L. et al. Extracting urban patterns from location-based social networks. In: *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks*. New York, NY, USA: ACM, 2011. (LBSN '11), p. 9–16. ISBN 978-1-4503-1033-8. Disponível em: <http://doi.acm.org/10.1145/2063212.2063226>. Citado na página 38.

GABRIELLI, L. et al. From tweets to semantic trajectories: mining anomalous urban mobility patterns. In: *Citizen in Sensor Networks*. [S.l.]: Springer, 2014. p. 26–35. Citado 2 vezes nas páginas 29 e 30.

GONG, V. X. *Exploring Human Activity Patterns Across Cities through Social Media Data*. Dissertação (Mestrado) — Faculty EEMCS, Delft University of Technology, Delft, the Netherlands, 2016. Citado na página 41.

GONZALEZ, M. C.; HIDALGO, C. A.; BARABASI, A.-L. Understanding individual human mobility patterns. *Nature*, Nature Publishing Group, v. 453, n. 7196, p. 779–782, 2008. Citado 3 vezes nas páginas 20, 57 e 58.

HAO, Q. et al. Equip tourists with knowledge mined from travelogues. In: *Proceedings of the 19th International Conference on World Wide Web*. New York, NY, USA: ACM, 2010. (WWW '10), p. 401–410. ISBN 978-1-60558-799-8. Disponível em: <http://doi.acm.org/10.1145/1772690.1772732>. Citado na página 18.

HASAN, S.; ZHAN, X.; UKKUSURI, S. V. Understanding urban human activity and mobility patterns using large-scale location-based data from online social media. In: *Proceedings of the 2Nd ACM SIGKDD International Workshop on Urban Computing*. New York, NY, USA: ACM, 2013. (UrbComp '13), p. 6:1–6:8. ISBN 978-1-4503-2331-4. Disponível em: <http://doi.acm.org/10.1145/2505821.2505823>. Citado 2 vezes nas páginas 37 e 57.

HAWELKA, B. et al. Geo-located twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science*, Taylor & Francis, v. 41, n. 3, p. 260–271, 2014. Citado na página 37.

HSIEH, H.-P.; LI, C.-T.; LIN, S.-D. Exploiting large-scale check-in data to recommend time-sensitive routes. In: *Proceedings of the ACM SIGKDD International Workshop on Urban Computing*. New York, NY, USA: ACM, 2012. (UrbComp '12), p. 55–62. ISBN 978-1-4503-1542-5. Disponível em: <http://doi.acm.org/10.1145/2346496.2346506>. Citado na página 18.

HUANG, Q.; CAO, G.; WANG, C. From where do tweets originate?: A gis approach for user location inference. In: *Proceedings of the 7th ACM SIGSPATIAL International Workshop on Location-Based Social Networks*. New York, NY, USA: ACM, 2014. (LBSN '14), p. 1–8. ISBN 978-1-4503-3140-1. Disponível em: <http://doi.acm.org/10.1145/2755492.2755494>. Citado na página 52.

JIANG, S. et al. A review of urban computing for mobile phone traces: Current methods, challenges and opportunities. In: *Proceedings of the 2Nd ACM SIGKDD International Workshop on Urban Computing*. New York, NY, USA: ACM, 2013. (UrbComp '13), p. 2:1–2:9. ISBN 978-1-4503-2331-4. Disponível em: <http://doi.acm.org/10.1145/2505821.2505828>. Citado na página 20.

JURDAK, R. et al. Understanding human mobility from twitter. *PloS one*, Public Library of Science, v. 10, n. 7, p. e0131469, 2015. Citado na página 37.

KAPLAN, A. M.; HAENLEIN, M. Users of the world, unite! the challenges and opportunities of social media. *Business Horizons*, v. 53, n. 1, p. 59 – 68, 2010. ISSN 0007-6813. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0007681309001232>. Citado na página 28.

KISILEVICH, S.; MANSMANN, F.; KEIM, D. P-dbscan: A density based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos. In: *Proceedings of the 1st International Conference and Exhibition on Computing for Geospatial Research & Application*. New York, NY, USA: ACM, 2010. (COM.Geo '10), p. 38:1–38:4. ISBN 978-1-4503-0031-5. Disponível em: <http://doi.acm.org/10.1145/1823854.1823897>. Citado na página 51.

LI, L.; GOODCHILD, M. F.; XU, B. Spatial, temporal, and socioeconomic patterns in the use of twitter and flickr. *cartography and geographic information science*, Taylor & Francis, v. 40, n. 2, p. 61–77, 2013. Citado na página 41.

LUO, F. et al. Explore spatiotemporal and demographic characteristics of human mobility via twitter: A case study of chicago. *Applied Geography*, v. 70, p. 11 – 25, 2016. ISSN 0143-6228. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0143622816300194>. Citado 4 vezes nas páginas 41, 52, 57 e 83.

MIAO, Z. et al. Cost-effective online trending topic detection and popularity prediction in microblogging. *ACM Trans. Inf. Syst.*, ACM, New York, NY, USA, v. 35, n. 3, p. 18:1–18:36, dez. 2016. ISSN 1046-8188. Disponível em: <http://doi.acm.org/10.1145/3001833>. Citado na página 29.

MILES, J.; SHEVLIN, M. *Applying regression and correlation: A guide for students and researchers*. [S.l.]: Sage, 2001. Citado na página 72.

MONTOLIU, R.; BLOM, J.; GATICA-PEREZ, D. Discovering places of interest in everyday life from smartphone data. *Multimedia tools and applications*, Springer, v. 62, n. 1, p. 179–207, 2013. Citado na página 51.

NAGHETTINI, M.; PINTO, É. J. d. A. *Hidrologia estatística*. [S.l.]: CPRM, 2007. Citado na página 32.

NETO, F. D. N.; BAPTISTA, C. de S.; CAMPELO, C. E. C. Prediction of destinations and routes in urban trips with automated identification of place types and stay points. *Revista Brasileira de Cartografia*, v. 68, n. 6, 2016. Citado 3 vezes nas páginas 29, 30 e 52.

NGUYEN, T.; SZYMANSKI, B. K. Using location-based social networks to validate human mobility and relationships models. In: IEEE. *Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on*. [S.l.], 2012. p. 1215–1221. Citado na página 39.

NOULAS, A. et al. A tale of many cities: universal patterns in human urban mobility. *PloS one*, Public Library of Science, v. 7, n. 5, p. e37027, 2012. Citado 2 vezes nas páginas 18 e 39.

OLIVEIRA, M. G. *Ontology-driven urban issues identification from social media.* Tese (Doutorado) — Federal University of Campina Grande, Brazil, 2017. Citado na página 49.

PALCHYKOV, V. et al. Inferring human mobility using communication patterns. *arXiv preprint arXiv:1404.7675*, v. 4, n. 6174, p. 6, 2014. Citado na página 20.

RHEE, I. et al. On the levy-walk nature of human mobility. *IEEE/ACM transactions on networking (TON)*, IEEE Press, v. 19, n. 3, p. 630–643, 2011. Citado na página 20.

SAIF, H. et al. Contextual semantics for sentiment analysis of twitter. *Information Processing & Management*, Elsevier, v. 52, n. 1, p. 5–19, 2016. Citado na página 29.

SAKAKI, T.; OKAZAKI, M.; MATSUO, Y. Earthquake shakes twitter users: Real-time event detection by social sensors. In: *Proceedings of the 19th International Conference on World Wide Web*. New York, NY, USA: ACM, 2010. (WWW '10), p. 851–860. ISBN 978-1-60558-799-8. Disponível em: <http://doi.acm.org/10.1145/1772690.1772777>. Citado 3 vezes nas páginas 29, 30 e 31.

SHIN, R. et al. On the levy-walk nature of human mobility: Do humans walk like monkeys? In: *Proc. IEEE INFOCOM*. [S.l.: s.n.], 2008. p. 924–932. Citado na página 58.

STEIGER, E. et al. Mining and correlating traffic events from human sensor observations with official transport data using self-organizing-maps. *Transportation Research Part C: Emerging Technologies*, v. 73, p. 91 – 104, 2016. ISSN 0968-090X. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0968090X16302030>. Citado na página 40.

STEIGER, E. et al. Twitter as an indicator for whereabouts of people? correlating twitter with {UK} census data. *Computers, Environment and Urban Systems*, v. 54, p. 255 – 265, 2015. ISSN 0198-9715. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0198971515300181>. Citado na página 41.

WAKAMIYA, S.; LEE, R.; SUMIYA, K. Crowd-based urban characterization: Extracting crowd behavioral patterns in urban areas from twitter. In: *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks*. New York, NY, USA: ACM, 2011. (LBSN '11), p. 77–84. ISBN 978-1-4503-1033-8. Disponível em: <http://doi.acm.org/10.1145/2063212.2063225>. Citado na página 36.

WEI, C.-P.; PIRAMUTHU, S.; SHAW, M. J. Knowledge discovery and data mining. In: _____. *Handbook on Knowledge Management: Knowledge Directions*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003. p. 157–189. ISBN 978-3-540-24748-7. Disponível em: <http://dx.doi.org/10.1007/978-3-540-24748-7_9>. Citado 3 vezes nas páginas 25, 26 e 27.

WILSON, T.; BELL, M. Comparative empirical evaluations of internal migration models in subnational population projections. *Journal of Population Research*, v. 21, n. 2, p. 127, 2004. ISSN 1835-9469. Disponível em: <http://dx.doi.org/10.1007/BF03031895>. Citado na página 18.

XIE, W. et al. Topicsketch: Real-time bursty topic detection from twitter. *IEEE Transactions on Knowledge and Data Engineering*, IEEE, v. 28, n. 8, p. 2216–2229, 2016. Citado na página 29.

YIN, H. et al. Joint modeling of users' interests and mobility patterns for point-of-interest recommendation. In: *Proceedings of the 23rd ACM International Conference on Multimedia*. New York, NY, USA: ACM, 2015. (MM '15), p. 819–822. ISBN 978-1-4503-3459-4. Disponível em: <http://doi.acm.org/10.1145/2733373.2806339>. Citado na página 38.

YUAN, Q. et al. Who, where, when and what: Discover spatio-temporal topics for twitter users. In: *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2013. (KDD '13), p. 605–613. ISBN 978-1-4503-2174-7. Disponível em: <http://doi.acm.org/10.1145/2487575.2487576>. Citado na página 36.

ZANDBERGEN, P. A. Accuracy of iphone locations: A comparison of assisted gps, wifi and cellular positioning. *Transactions in GIS*, Wiley Online Library, v. 13, n. s1, p. 5–25, 2009. Citado na página 36.

ZAR, J. H. Significance testing of the spearman rank correlation coefficient. *Journal of the American Statistical Association*, Taylor & Francis Group, v. 67, n. 339, p. 578–580, 1972. Citado na página 32.

ZHANG, C. et al. A hybrid term–term relations analysis approach for topic detection. *Knowledge-Based Systems*, v. 93, p. 109 – 120, 2016. ISSN 0950-7051. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0950705115004335>. Citado na página 29.

ZHANG, Y. et al. Towards a temporal network analysis of interactive wifi users. *EPL (Europhysics Letters)*, IOP Publishing, v. 98, n. 6, p. 68002, 2012. Citado na página 20.

ZHAO, K. et al. Explaining the power-law distribution of human mobility through transportation modality decomposition. *arXiv preprint arXiv:1408.4910*, v. 5, n. 9136, p. 21, 2014. Citado na página 20.

ZHENG, V. W. et al. Collaborative location and activity recommendations with gps history data. In: *Proceedings of the 19th International Conference on World Wide Web*. New York, NY, USA: ACM, 2010. (WWW '10), p. 1029–1038. ISBN 978-1-60558-799-8. Disponível em: <http://doi.acm.org/10.1145/1772690.1772795>. Citado na página 18.