Capítulo 3

Dados geoespaciais: Conceitos e técnicas para coleta, armazenamento, tratamento e visualização

Augusto Cesar Souza Araujo Domingues, Fabrício Aguiar Silva, Leonardo Júnio Alves dos Santos, Raissa Polyanna Papini de Melo Souza, Gabriel Teixeira Pinto Coimbra e Antonio Alfredo Ferreira Loureiro

Abstract

In recent years, there has been a significant increase in the availability of data from mobile objects (e.g., vehicles, people, drones and others) that have geospatial information (time and space). From these data, companies, public institutions and the scientific community are working to extract useful knowledge, aiming to know and understand the behavior of urban mobility, thus being able to improve services and propose solutions that advance the state of the art in areas related to mobility, Internet of Things, urban computing, among others. The goal of this chapter is to present the main theoretical and technical concepts related to the treatment of geospatial data including its collection, storage, treatment and visualization. For this, we intend to discuss the theoretical concepts associated with these issues, which are relevant to the process of manipulating geospatial data, as well as the technical content with a focus on the main existing tools and libraries.

Resumo

Nos últimos anos, houve um aumento significativo na disponibilidade de dados oriundos de objetos móveis (e.g., veículos, pessoas, drones e outros), os quais geram informações geoespaciais (tempo e espaço). A partir desses dados, empresas, instituições públicas e a comunidade científica estão atuando na pesquisa e na proposta de soluções capazes de extrair conhecimento útil, com o intuito de conhecer e entender o comportamento da mobilidade urbana em seus diversos aspectos. Essas soluções podem melhorar os serviços ofertados pelas empresas e gerar novos insights para captação de novas receitas, além de avançar o estado da arte em áreas relacionadas com mobilidade, Internet das Coisas, computação urbana, dentre outras. O objetivo deste capítulo é alinhar a teoria e a prática, apresentando os principais conceitos e técnicas associadas ao tratamento

e manipulação de dados geoespaciais, o que inclui as fases de coleta, armazenamento, tratamento e visualização, com a utilização das principais ferramentas e bibliotecas existentes.

3.1. Introdução

A mobilidade é um dos fatores mais importantes no atual cenário tecnológico, político e econômico mundial. É com base nela que políticas e serviços, antes criados de maneira genérica, estão sendo aperfeiçoados e personalizados para garantir uma melhor experiência [Hess et al. 2015]. Quando falamos de mobilidade humana especificamente, envolvemos questões como predição de fluxo de trânsito, modelos de contágio, otimização de recursos de rede, planejamento urbano, análise do comportamento social e até estudos sobre fluxos migratórios [Barbosa et al. 2018]. Para tratar dessas questões, duas frentes de trabalho são comumente empregadas. A primeira, mais tradicional, ocorre através da construção de modelos matemáticos e estatísticos que permitem derivar o comportamento de mobilidade do objeto estudado com certo grau de realismo. Por outro lado, devido à crescente coleta de dados geoespaciais através de interfaces como dispositivos móveis e redes sociais, tornou-se notável, nos últimos anos, uma segunda frente de trabalho baseada no estudo da mobilidade através da análise de dados históricos, isto é, coleções de registros de mobilidade gerados pelos objetos em estudo. Essas coleções – frequentemente chamadas de *traces de mobilidade* – permitem a construção de modelos com alto grau de realismo sem que seja necessário algum conhecimento prévio dos objetos.

Diante dos avanços da análise de dados históricos, a cada dia novas empresas utilizam dados geoespaciais disponíveis para criarem produtos e serviços que se beneficiam das informações de mobilidade para oferecer experiências mais personalizadas. Nesse cenário, podemos citar como exemplo as redes sociais baseadas em localização (*Location-Based Social Networks* ou LBSNs), os serviços de compartilhamento de veículos e os seguros veiculares baseados em comportamento. Adicionalmente, o setor público passou a investir na captação de dados geoespaciais (posição em tempo real de veículos de patrimônio público, localização de imóveis e infraestrutura em geral, entre outros), a fim de fornecer à população uma interface de transparência na gestão de seus recursos. Servem de exemplo os portais de dados das cidades de Nova Iorque[1], Chicago[2] e Seattle[3], que oferecem inúmeras coleções de livre acesso. Finalmente, a comunidade acadêmica tem investigado essas bases de dados de mobilidade para gerar novos conhecimentos que auxiliam na tomada de decisão em diferentes áreas.

Para ilustrar o crescente interesse no estudo de dados geoespaciais, a Figura 3.1 apresenta um levantamento dos últimos vinte anos do número de artigos publicados por ano em conferências e periódicos nacionais e internacionais que envolvem o tópico de dados geoespaciais[4]. Como pode ser observado, em 2010 o número de artigos publicados foi dez vezes maior quando comparado ao ano de 2000. Ademais, o total de trabalhos publicados no último ano (2019) foi mais de duas vezes maior que o número registrado em 2010. Nesse cenário, todos os fatores indicam que essa curva continuará a crescer a cada

[1] `https://opendata.cityofnewyork.us/`
[2] `https://data.cityofchicago.org/`
[3] `https://data.seattle.gov/`
[4] `https://www.webofscience.com`

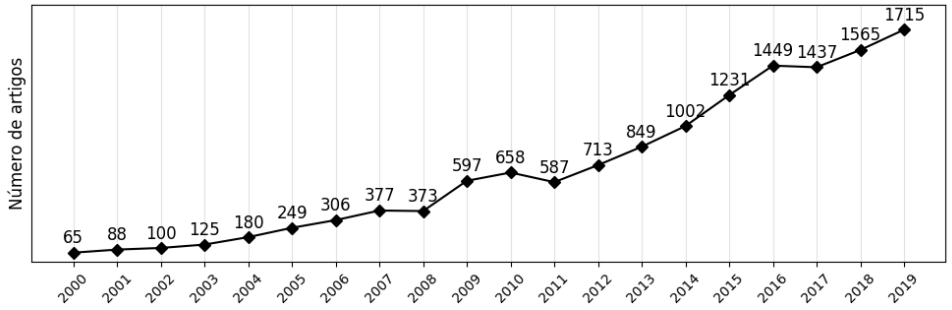


Figura 3.1. Artigos publicados por ano com o tópico "geospatial data" (Fonte: *Web of Science*).

ano, devido à popularização e à disponibilidade cada vez maior de dados geoespaciais.

A demanda por novos estudos introduz a necessidade do desenvolvimento de novas técnicas e ferramentas para a manipulação de dados geoespaciais. A seguir, são destacadas questões específicas para esse tipo de dado a serem tratadas dentre as etapas de manipulação e tratamento de dados:

- Coleta: No que compete ao processo de coleta dos dados geoespaciais, é necessário definir as estratégias a serem adotadas, considerando os diferentes tipos de sensores de localização existentes e os problemas inerentes desses sistemas, como a acurácia das coletas, a privacidade dos usuários, a frequência de amostragem e a escala dos dados;

- Armazenamento: Devido às particularidades dos dados geoespaciais, é preciso considerar estratégias eficientes de seu armazenamento, que utilizem sistemas com indexação e compactação específicas para esse tipo de dado;

- Preparação e extração: Para garantir a qualidade dos resultados, é preciso primeiro garantir a qualidade dos dados que serão usados como entrada. Para isso, faz-se necessário que os mesmos sejam preparados, o que inclui etapas de amostragem, limpeza, filtragem e agregação.

- Visualização: Durante as etapas da análise de dados geoespaciais, uma tarefa frequente é a visualização de resultados parciais. Por sua vez, é fundamental apresentar os resultados finais de maneira clara e concisa. Dito isso, é preciso escolher as formas mais adequadas para a exibição de dados geoespaciais.

É válido ressaltar que ainda não há um consenso em relação à metodologia adequada para a análise de dados geoespaciais, em termos dos tópicos apresentados acima. Isso ocorre devido à escassez de referências que apresentem, de forma clara, os conceitos básicos e técnicas aplicadas na análise desse tipo de dado. Assim, visando preencher essa lacuna, este trabalho apresenta de maneira extensa um material que contempla a teoria relacionada às etapas do processamento de dados geoespaciais – da coleta à aplicação. Além disso, introduzimos as principais ferramentas e bibliotecas utilizadas para a análise desse tipo de dado na prática. Esperamos que ao fim deste capítulo o leitor seja capaz de produzir resultados relevantes a partir da análise de dados geoespaciais, adotando as práticas mais

apropriadas e selecionando as ferramentas e algoritmos mais adequados ao conjunto de dados escolhido.

3.1.1. Por que dados geoespaciais?

A seguir, apresentamos algumas das principais aplicações para as quais o uso de dados geoespaciais é benéfico, com o intuito de motivar o leitor a contemplar a importância desse tipo de dado e também do seu uso de forma adequada. Para cada aplicação, destacamos como os dados podem ser coletados e aplicados como fonte de informação relevante.

3.1.1.1. Mobilidade Urbana

Entender e modelar o comportamento urbano de pessoas, veículos e outros objetos móveis é um dos pilares da computação urbana [Zheng et al. 2014]. Através do conhecimento gerado, podemos planejar melhor o futuro de centros urbanos e, consequentemente, melhorar a qualidade de vida de seus habitantes. Ademais, esse conhecimento nos permite compreender como ocorrem as relações sociais e como elas interferem no nosso dia a dia, especialmente do ponto de vista de acesso a recursos computacionais [Wang e Song 2015]. Dessa maneira, dados geoespaciais podem prover informações sobre as dinâmicas de mobilidade de milhões de pessoas, sendo mais precisos e baratos de se obter quando comparados a estratégias convencionais de coletas de dados, tais como pesquisas de campo e contadores de tráfego [Naboulsi et al. 2016].

Do ponto de vista da aplicação de dados geoespaciais no estudo da mobilidade urbana, destacam-se as fontes de dados de grande escala populacional, como as redes sociais baseadas em localização (LBSN), os serviços de telefonia e as fontes de dados públicos. Os dados de LBSN, como *Twitter* e *Foursquare*, podem ser usados para o sensoriamento coletivo de tráfego [Santos et al. 2018] e para a identificação de pontos de interesse [Gu et al. 2016]. Já as operadoras de serviços de telefonia disponibilizam registros de chamadas e de dados móveis, que podem ser aplicados no planejamento e na alocação de recursos de rede, permitindo, por exemplo, a estimativa da demanda em grandes eventos [Gao 2015, Marques-Neto et al. 2018]. Por fim, temos os dados de serviços públicos, como registros de viagens de táxi e de ônibus, que possibilitam análises dos fluxos urbanos e de demanda de serviços [Castro et al. 2012].

3.1.1.2. Internet dos Drones

De acordo com [Motlagh et al. 2016], a expectativa é que em alguns anos milhões de drones estejam disponíveis para atuar em diversos setores da economia, realizando atividades como entregas de encomendas, mapeamento e vigilância de áreas de difícil alcance, agricultura e até na atuação direta em combates. Para que isso ocorra, é preciso facilitar a mobilidade e a comunicação entre os veículos aéreos não-tripulados, por meio do desenvolvimento de algoritmos de roteamento de mensagens e métodos de orquestração. Essas tecnologias farão uso dos sensores geoespaciais desses veículos, tais como sensores GPS, câmeras de alta definição e sensores *bluetooth*, para permitir a criação de grupos de drones e suas respectivas movimentações.

3.1.1.3. Redes Veiculares

As VANETs (Vehicular Ad hoc Networks) são redes que permitem a comunicação entre veículos e entre unidades auxiliares instaladas nas vias, a fim de prover o funcionamento de serviços como alertas de tráfego e acidentes, compartilhamento de multimídia, dentre outros. O seu objetivo é tornar a mobilidade mais agradável e segura para motoristas, passageiros e pedestres. Dessa maneira, é necessário modelar a mobilidade veicular para que as aplicações, os serviços e os protocolos da rede usufruam desta informação e consigam se adaptar ao comportamento dos veículos. Assim, fontes de dados como *traces* de mobilidade de táxis, ônibus e veículos particulares são de suma importância para o desenvolvimento dessas tecnologias, sendo utilizados tanto para as análises de comportamento, gerando modelos de mobilidade, quanto para a validação de algoritmos e protocolos propostos para serem usados em ambientes urbanos.

3.1.1.4. Comunicação Oportunista

O roteamento de mensagens de forma oportunística é um paradigma de comunicação entre dispositivos integrantes de uma rede, que permite a transmissão de mensagens entre dois ou mais componentes que se encontram por um determinado intervalo de tempo. Tipicamente, esses contatos ocorrem de forma intermitente, devido a fatores contextuais como a velocidade de veículos, as rotas seguidas, o raio de comunicação para haver uma transmissão e aspectos sociais da mobilidade. Tais fatores podem ser usados para definir estratégias de roteamento, reduzindo assim o número de mensagens enviadas sem que haja perda na qualidade do serviço. O uso de dados geoespaciais permite o desenvolvimento e a validação dessas estratégias em diferentes ambientes e situações, sem que seja necessário a realização de experimentos reais, o que atualmente é, em geral, inviável.

3.1.1.5. Controle de Epidemias e Modelos de Contágio

Dados geoespaciais são capazes de capturar o comportamento de mobilidade humana e suas características, como pontos de interesse, interações sociais, padrões de mobilidade de pessoas e os fluxos de mobilidade existentes. Esses fatores tornam possível o uso desses dados para a construção e aperfeiçoamento de modelos de contágio, que por sua vez permitem estimar os efeitos de doenças infecciosas sobre uma população. Podemos citar como exemplo os modelos Susceptível-Infectado-Recuperado (SIR) e Susceptível-Infectado-Susceptível (SIS), que podem ser usados para mapear o comportamento de doenças como rubéola e gripe, respectivamente. Ao empregarmos as informações obtidas dos dados geoespaciais aos modelos de contágio, podemos estimar as taxas de infecção e transmissão em uma dada população, auxiliando no desenvolvimento de ações e medidas preventivas. Outra aplicação no controle de epidemias está sendo adotada em vários lugares do mundo, em que dados geoespaciais têm sido usados para realizar o monitoramento populacional e o controle da formação de aglomerações durante o período de confinamento devido à pandemia do novo coronavírus (SARS-CoV2).

3.1.1.6. Segurança

Por fim, a análise de dados geoespaciais – gerados por meio de câmeras, sensores de presença, redes celulares e redes sem fio – pode auxiliar na proteção e na segurança de usuários. Nas redes veiculares, por exemplo, a análise de dados relacionados ao próprio veículo e os que estão na vizinhança, às condições viárias e do ambiente onde o veículo se encontra pode auxiliar o condutor na prevenção de acidentes, provendo sistemas de condução assistida ou autônoma, frenagem automática e alertas de trânsito e colisão, dentre outras possibilidades. Quanto aos aspectos de segurança pública e privada, os dados geoespaciais fornecem uma cobertura espacial que, juntamente com o sensoriamento coletivo, permitem que cada usuário na rede contribua para a segurança de todos. Adicionalmente, essa análise contribui para o mapeamento do comportamento de entidades suspeitas, permitindo e facilitando a predição e resolução de crimes.

3.2. Conceitos Fundamentais

Nesta seção, serão apresentados os principais conceitos relacionados a dados geoespaciais, essenciais em todas as etapas da análise desses dados. Primeiramente, destacamos as características geográficas da Terra e como estas podem afetar diversas operações com dados geoespaciais (Seção 3.2.1). Em seguida, considerando um contexto mais amplo, discutimos o que são sistemas de referências (Seção 3.2.2). A manipulação de dados geoespaciais através de projeções espaciais é discutida na Seção 3.2.3. Por fim, abordamos a manipulação numérica de coordenadas, destacando as particularidades existentes no uso de operações matemáticas para o tratamento de dados geoespaciais (Seção 3.2.4).

Independentemente de como os dados geoespaciais são obtidos e registrados para criarmos uma base de dados, ao coletarmos esses dados continuamente precisamos assinalar o instante no tempo desse registro. Assim, dados geoespaciais têm também a dimensão temporal.

3.2.1. Geografia e suas propriedades

As ciências geodésicas são responsáveis pelo estudo da forma e da superfície da terra, considerando suas imperfeições e os objetos – naturais ou artificiais – existentes sobre (e sob) ela. Ela trata do levantamento das informações e da definição de representações e medidas. Segundo [Bolstad 2016], para usarmos de maneira efetiva dados geoespaciais e os sistemas que derivam deles, é preciso estabelecer um entendimento claro de como os sistemas de coordenadas são definidos para a Terra, como essas coordenadas são medidas sobre a superfície curva da mesma, e por fim como são convertidas em diversas projeções para seu uso, seja manual ou digital. Se esses fatores não forem levados em consideração, os dados geoespaciais coletados serão imprecisos e, consequentemente, as operações realizadas sobre eles poderão gerar resultados com erros. Enquanto tal imprecisão possa parecer pequena e irrelevante para alguns casos, aplicações de alto risco como o cálculo da trajetória de mísseis não podem permiti-la. Podemos definir dois fatores principais a serem considerados quanto à geografia da terra: o seu formato e a imprecisão das medidas.

São três os modelos comumente usados para representar a superfície terrestre: projeção plana, formato esférico e formato elíptico. Apesar de permitirem e facilitarem a

visualização de mapas em superfícies 2D, projeções planas da Terra distorcem sua geometria curva. Um exemplo disso é a projeção de uma linha reta entre dois pontos quaisquer em um mapa plano, resultando na omissão ao leitor da curvatura existente na superfície terrestre entre os dois extremos da linha. Por sua vez, o modelo de formato esférico assume a Terra como uma esfera, isto é, todos os pontos de sua superfície estão a uma mesma distância do seu centro. Apesar de eliminar as limitações existentes na projeção plana, o modelo esférico ignora o achatamento da Terra em seus polos, podendo levar a medidas imprecisas sobre essas regiões. Finalmente, o modelo elíptico adota essa propriedade, sendo o mais semelhante à geometria da Terra. Na medida em que se tornam mais precisos, os modelos também se tornam mais complexos, requerendo medições e cálculos mais avançados. Isso pode afetar a eficiência da solução proposta quanto ao tempo de execução e ao espaço (memória), o que deve ser levado em consideração durante a escolha do melhor modelo.

É válido ressaltar que os modelos existentes são representações simplificadas do formato real da terra e, portanto, sempre terão imperfeições. Capturar todas as variações geográficas – naturais e artificiais – da superfície terrestre em um dado instante não é viável, mesmo se considerarmos um modelo perfeito do formato da Terra. Medições imprecisas ocorrerão devido aos limites dos sensores utilizados para realizar suas coletas. Sensores GPS de *smartphones*, por exemplo, possuem acurácia de aproximadamente 5 m, podendo diminuir na presença de obstáculos como montanhas, túneis e prédios.

Na prática, escolhemos o modelo de acordo com a resolução espacial do problema a ser tratado. Visualizações em pequena escala costumam ser feitas através de projeções planas, devido à baixa interferência da curvatura da Terra sobre as mesmas. Já os sistemas de recomendação de rotas, que realizam cálculos de distâncias precisas, utilizam de modelos esféricos ou elípticos simples. Por fim, sistemas bélicos que envolvem o cálculo de trajetórias de projéteis necessitam de modelos elípticos que considerem a geografia local, ou seja, a presença de vales, montanhas, prédios, entre outros.

3.2.2. Sistemas de coordenadas

Sistemas de coordenadas usam coordenadas para determinar a posição de objetos em um espaço. Esse espaço pode ser composto de uma ou mais dimensões, e cada dimensão pode conter propriedades particulares como limites inferiores e superiores, notações e escalas. Assim, para um sistema de coordenadas cartesianas de duas dimensões, podemos definir a localização de um objeto sobre o plano cartesiano pelo par de coordenadas $P = (x, y)$, onde cada valor representa a posição em uma das dimensões. De maneira análoga, sistemas de coordenadas geográficas permitem representar a localização de objetos sobre a superfície terrestre através de coordenadas geográficas. Essas podem ter duas ou mais dimensões e assumir diferentes modelos para o formato da Terra. A seguir, discutiremos sobre o sistema de coordenadas geográficas e suas propriedades, e apresentaremos outros sistemas de referências frequentemente encontrados em dados geoespaciais, como registros de telefonia celular, registros de encontros e registros de *check-in*.

O modelo mais usado de coordenadas geográficas se baseia em um sistema de coordenadas esféricas para permitir a localização de objetos em uma superfície que se assemelhe ao formato da Terra. Esse sistema utiliza de dois ângulos de rotação para

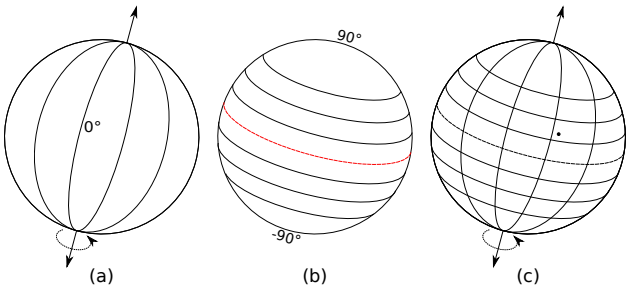


Figura 3.2. Sistema de coordenadas geográficas usado para localizar objetos sobre a Terra. (a) Eixo longitudinal varia de -180° a 180°, sendo o ponto zero o meridiano de Greenwich. As linhas representam pontos de longitude constante e são denominadas de meridianos. (b) Eixo latitudinal varia de -90° a 90°, sendo o ponto zero a linha do Equador (em vermelho). As linhas representam pontos de latitude constante e são denominadas de paralelos. (c) A combinação de uma longitude com uma latitude fornece a localização precisa de um objeto.

especificar posições na superfície modelada. O primeiro ângulo de rotação, denominado longitude (Figura 3.2a), é calculado ao redor do eixo imaginário sobre o qual a Terra realiza o seu movimento de rotação. Esse eixo atravessa o centro da Terra e tem como extremidades os Polos Norte e Sul. A variação posicional sobre o eixo é medida em graus, com a posição zero (0°) localizada sobre uma linha imaginária (um meridiano) que passa próximo ao Observatório Real de Greenwich, na Inglaterra. A variação é positiva no sentido leste e negativa no sentido oeste, atingindo o valor máximo de 180° (ou -180°) exatamente no ponto oposto à posição zero sobre a superfície terrestre. O segundo ângulo de rotação, denominado latitude (Figura 3.2b), é calculado sobre a linha do Equador, que representa a metade da distância entre o Polo Norte e o Polo Sul. Sua posição zero (0°) está localizada exatamente sobre a linha do Equador, com variações no sentido Norte tendo sinal positivo e variações no sentido Sul tendo sinal negativo, atingindo valores máximos nos Polos Norte e Sul de 90° e -90°, respectivamente. Desta forma, podemos definir a posição de um objeto sobre a Terra através de um par de ângulos latitude e longitude (Figura 3.2c). Por sua vez, cada grau pode ser dividido em 60 minutos e cada minuto em 60 segundos, o permite às coordenadas geográficas de latitude e longitude especificarem a localização de um objeto com precisão abaixo de 1 m. Por convenção, os ângulos sempre são especificados na ordem (*latitude*, *longitude*).

3.2.2.1. Registros de Telefonia Celular

Os registros de telefonia celular (*Call Detail Records – CDR*) são coletados pelas operadoras de redes de telefonia celular para realizar o faturamento dos planos de seus clientes. A cada ligação realizada, mensagem *SMS* enviada ou sessão de dados terminada, um registro é produzido, contendo informações como identificação do remetente e do destinatário, horário, duração e localização. Quanto à localização, os registros possuem acurácia limitada, pois a posição reportada indica somente em qual célula da rede móvel o usuário se encontra. As células na rede cobrem regiões de milhares de metros quadrados, e não é possível estimar a localização real do usuário na célula. Outro fator é a granularidade temporal dos registros, que só são criados quando um usuário realiza uma das ações

apresentadas acima. Sendo assim, usuários com perfil reduzido de uso terão grandes lacunas entre seus acessos, ou seja, intervalos de tempo onde não se têm informação sobre o comportamento dos mesmos.

3.2.2.2. Registros de *Check-In*

Os registros de *check-in* (visitas) representam a presença de um usuário em algum local de interesse (*Point of Interest – PoI*) em um determinado instante. Os PoIs podem ser locais públicos, como restaurantes, lojas e mercados, como também locais privados e específicos de cada usuário, como casa e local de trabalho. Os registros de *check-in* são obtidos através de LBSNs como *Facebook*, *Twitter* e *Foursquare* que capturam as informações sobre o usuário, o local visitado e o horário da visita, podendo ou não conter as coordenadas geográficas do local. Assim como os registros de telefonia celular, os registros de *check-in* possuem granularidade temporal grossa, pois as visitas feitas por um usuário durante o dia só são registradas se o mesmo inseri-las voluntariamente através de suas LBSNs, o que nem sempre ocorre devido a questões como privacidade (o usuário não deseja que outros saibam o local onde se encontra) e segurança (o usuário teme que reportar sua localidade irá colocar em risco sua segurança).

3.2.2.3. Registros de Encontros

Os registros de encontros, também chamados de registros de contatos, representam a ocorrência de um encontro entre duas ou mais entidades de acordo com uma regra pré-estabelecida. Essa regra varia conforme as características dos dados analisados e o objetivo da análise a ser feita. A distância entre as localizações atuais das entidades, a presença simultânea em um mesmo local e a conexão simultânea em uma mesma rede são critérios comumente adotados para caracterizar os encontros. De maneira similar, são definidos critérios para o encerramento de um encontro. O encontro se inicia no momento no qual a regra de ocorrência passa a ser atendida pelas entidades, e termina quando a regra de encerramento é atendida.

Os registros de encontros podem ser gerados tanto a partir de outros tipos de registros (coordenadas GPS, de telefonia celular e de *check-ins*) quanto a partir da coleta por proximidade (*bluetooth*, *beacons*). Assim, são aplicáveis em diversos tipos de análises, como a validação de algoritmos de transmissão, modelos de contágio, estudos de interações sociais, entre outros.

3.2.3. Projeções Espaciais

Dados geoespaciais nos fornecem a localização precisa de objetos sobre a terra através de ângulos de latitude e longitude. Porém, às vezes precisamos representar esta posição sobre superfícies com formatos diferentes, como um mapa plano. Mapas cobrem superfícies maiores ao mesmo tempo, facilitam a visualização em papel e *displays*, e sua produção é simples. Por outro lado, é impossível aplicarmos diretamente a posição de objetos em uma superfície esférica sobre uma superfície plana. Para isso, utilizamos de projeções

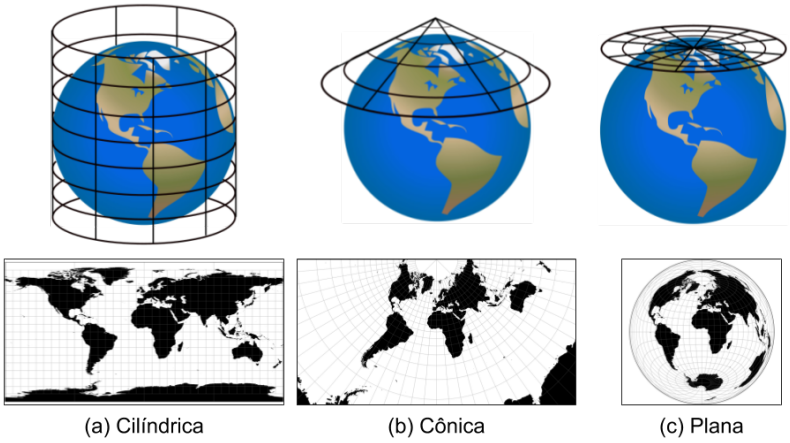


Figura 3.3. Tipos de projeções quanto a geometria de conversão.

espaciais, que realizam, através de fórmulas matemáticas, a renderização das localizações sobre a superfície original para a nova superfície.

Existem diversas projeções diferentes para o globo terrestre e, apesar de terem como fator comum a representação em uma superfície plana, elas variam quanto ao tipo e quanto às suas propriedades. O tipo da projeção se refere à forma geométrica utilizada para converter o globo em uma superfície plana (Figura 3.3). Essa forma pode ser cilíndrica, cônica, plana, ou alguma combinação dessas. Quanto às propriedades, elas representam as características que a projeção preserva em relação à superfície real da Terra, podendo ser conformal (preserva ângulos e formatos), equivalente (mantém as medidas de áreas reais), de compromisso (um meio termo entre conformal e equivalente) e equidistante (preserva a distância real entre pontos no mapa). A Tabela 3.3 apresenta a comparação das principais projeções encontradas na literatura quanto ao seu tipo e suas propriedades.

Diferentes projeções distorcem o globo terrestre de diferentes maneiras. Para se ajustar a um mapa plano é necessário criar distorções que podem comprimir ou alongar regiões do mapa. De fato, distorções do globo são inevitáveis em projeções planas. É o caso da projeção de *Mercator*, que para projetar os continentes em seu formato correto distorce as regiões próximas aos polos, que apresentam tamanho muito superior às suas áreas reais, o que pode gerar inconsistência entre as visualizações e os resultados numéricos obtidos das análises. Variante da projeção de *Mercator*, a projeção UTM (*Universal Mercator Transverse*) preserva os ângulos e formatos de regiões, porém ao custo da distorção de distâncias e áreas. Por outro lado, a projeção de *Gall-Peters* apresenta as superfícies com proporção exata a suas áreas, ao custo de distorcer o formato das mesmas. Por fim, a projeção *equidistante* preserva a distância real entre dois pontos quaisquer sobre a superfície, ao custo de distorções menores no formato e na área das regiões.

3.2.4. Manipulação Númerica de Coordenadas

Por fim, é necessário discutir a manipulação numérica de dados geoespaciais, que apesar de serem representados como números, não possuem o mesmo significado. Em razão

Tabela 3.1. Comparação das principais projeções encontradas na literatura quanto ao seu tipo e suas propriedades

Projeção	Tipo	Propriedades
Mercator	Cilíndrica	Conformal
UTM	Cilíndrica	Conformal
Gall-Peters	Cilíndrica	Área igual
Equidistante	Cilíndrica	Equidistante
Equidistance Cônica	Cônica	Equidistante
Azimuthal Equidistante	Plana	Equidistante

disso, diversas operações precisam ser modificadas para se adequarem à representação dos dados geoespaciais. O cálculo da média de coordenadas, por exemplo, pode resultar em valores inesperados devido à forma com a qual as coordenadas são distribuídas sobre o globo. Considere como exemplo a média de dois pontos localizados sobre os ângulos de longitude 179° e -179°. Ao realizarmos a operação, obtemos o ângulo médio de longitude de 0°, porém o resultado correto é 180°, devido ao encontro dos extremos longitudinais.

Além disso, é necessário considerar as diferentes formas de representação de coordenadas latitude e longitude – as variações podem ser representadas em graus, minutos e segundos ou em decimais – para realizar cálculos aritméticos e arredondamentos. Enquanto que os intervalos em minutos variam de 0 a 60 e os intervalos em segundos de 0 a 3600, os intervalos em decimais seguem a variação tradicional dos números reais.

Podemos citar também o cálculo do centro de polígonos formados por dados geoespaciais. Se utilizamos da média dos pontos que representam os extremos do polígono, o resultado poderá tender para alguma região na qual existam muitos pontos limítrofes. Assim, é necessário calcular o centroide do polígono, i.e., o seu centro de massa, que é constante para um mesmo formato e não depende da concentração dos pontos.

3.3. Coleta

Dados geoespaciais representam uma visão simplificada da relação de uma ou mais entidades físicas, como pessoas e veículos, em localidades, como rodoviais, cidades, coordenadas geográficas, pontos de interesse, entre outros. Ao capturarmos um *check-in* de um usuário de uma LBSN em um restaurante, extraímos as informações necessárias para representar esse evento como um dado geoespacial, como horário em que o *check-in* foi feito, nome e coordenadas geográficas do local (quando disponíveis). A definição das informações a serem coletadas deve ocorrer de acordo com a análise a ser feita, considerando também as limitações da tecnologia utilizada para realizar a coleta. Adicionalmente, questões como a privacidade dos usuários sensoreados e a ética das análises feitas devem ser levadas em consideração.

De forma geral, existem duas abordagens para a coleta de dados geoespaciais de entidades. A primeira delas ocorre através de medições locais e manuais, se deslocando para a localização da entidade a ser sensoreada e realizando as medições com o uso de ferramentas manuais e sensores, derivando dados geoespaciais precisos e ricos em contexto.

No entanto, esse método possui algumas ressalvas. Primeiramente, as localizações das entidades podem ser de difícil acesso, como regiões remotas ou de perigo elevado, colocando em risco a segurança do indivíduo responsável pelo mapeamento. Adicionalmente, a medição local pode violar a privacidade da entidade sensoreada ao realizar coletas em locais como sua casa ou local de trabalho. Aliam-se a esses fatores o alto custo de realizar as coletas, devido aos deslocamentos e ao tempo necessário, e a escalabilidade, dado a necessidade de realizar a coleta da localização de entidades em larga escala de maneira simultânea. O segundo método ocorre através de medições automáticas e é o mais utilizado devido a sua escalabilidade. Nele, as posições das entidades são obtidas através do uso de sensores e armazenadas para acesso posterior. Os sensores podem estar de posse das entidades sensoreadas, como *smartphones*, ou não, como satélites de captura de imagem e *drones*.

Nesta seção, discutimos sobre o processo de coleta de dados geoespaciais, fornecendo ao leitor os conceitos e ferramentas necessárias para realizá-lo. As fontes de dados existentes e suas propriedades são apresentadas na Seção 3.3.1 e, em seguida, discutimos diferentes aspectos relacionados à qualidade da coleta na Seção 3.3.2.

3.3.1. Fontes de dados

Coletar dados geoespaciais é uma atividade complexa e de alto custo. Portanto, é preciso ter um escopo bem definido para obter os resultados desejados sem que ocorram custos desnecessários com a coleta e processamentos posteriores. Uma das etapas da coleta é a escolha da fonte de dados, que deve considerar o compromisso entre a qualidade dos dados obtidos e o custo de aplicação da tecnologia de sensoriamento. Podemos considerar um exemplo em que se deseja mapear os pontos de interesse de uma região da cidade. Neste caso, dados de redes sem fio não possuem acurácia semelhante a de dados GPS, porém dados de endereços postais coletados através de LBSNs podem obter resultados semelhantes aos dados GPS e com custo inferior de coleta. Estas suposições também são válidas para o uso de dados geoespaciais coletados por terceiros que podem apresentar custos de aquisição.

3.3.1.1. Sensores de Sistemas de Navegação por Satélite

Sistemas de Navegação por Satélite (*Global Navigation Satellite Systems* – GNSS) são tecnologias baseadas em satélite que fornecem informação precisa sobre a localização de objetos sobre a superfície terreste, através do uso de receptores. Para isso, é necessário o uso de um receptor, que capta os sinais dos satélites para calcular a sua posição em ângulos de latitude e longitude. Esses sistemas são robustos, sendo capazes de operar ininterruptamente, independente de condições climáticas e em qualquer lugar na superfície da Terra, em ambientes externos. O GNSS mais utilizado na atualidade é o GPS (*Global Positioning System*, ou Sistema de Posicionamento Global), administrado pelo governo dos Estados Unidos. Além dele, está em operação o GLONASS (Rússia), e estão em desenvolvimento o Galileo (União Européia) e o Beidou (China). A seguir, referências feitas no texto ao Sistema GPS podem ser consideradas para qualquer GNSS genérico.

Os receptores GPS (ou sensores GPS) são classificados quanto à precisão de suas

medições em três categorias: geodésico, topográfico e de navegação. Enquanto os receptores geodésicos e topográficos possuem precisão em milímetros e centímetros, respectivamente, os receptores de navegação possuem precisão em metros, porém apresentam custo reduzido, o que os tornam preferenciais para o uso em *smartphones* e computadores portáteis. Para obter sua posição, um receptor capta o sinal de três satélites, obtendo também o tempo real ao captar o sinal de mais um satélite, totalizando quatro.

Por um lado, os sensores GPS são dispositivos de coleta de dados geoespaciais de melhor desempenho, obtendo localizações acuradas com alta frequência. Adicionalmente, estão presentes em grande parte dos dispositivos móveis encontrados na atualidade, como *smartphones*. Por outro lado, questões como privacidade dos usuários, eficiência energética dos sensores e perda de sinal dos satélites devem ser levadas em consideração. Em primeiro lugar, a acurácia e a frequência do sensoriamento da posição de um usuário permitem a obtenção de informações pessoais como a localização de sua casa e local de trabalho [Kang et al. 2004], como também de seus horários [Gu et al. 2016]. Em seguida, deve-se considerar o impacto do consumo elevado de energia dos sensores GPS em dispositivos móveis que têm fontes de energia limitadas, o que restringe a frequência de sua atualização. Além disso, a existência de barreiras como túneis, montanhas e *canyons* urbanos interrompe a captação de sinal dos satélites por parte dos sensores, gerando lacunas espaciais e temporais no sensoriamento [Silva et al. 2015]. Finalmente, erros de posição devido à precisão dos sensores causam inconsistências nas análises de trajetórias e pontos de interesse, demandando etapas de calibragem e limpeza dos dados antes de realizar as análises [Celes et al. 2017].

3.3.1.2. Redes sem fio

As redes de telefonia móvel e os pontos de acesso à internet via *Wi-Fi* compreendem o conjunto de redes sem fio. Capazes de serem usadas em dispositivos móveis, computadores e até veículos, as redes sem fio podem ser usadas como fontes de baixo custo de dados geoespaciais. As redes de telefonia móvel obtêm informações de localização de um usuário através das estações bases mais próximas, mesmo que o usuário não esteja realizando um acesso no momento. A posição reportada corresponde ao raio de alcance da torre contactada e, no caso onde duas ou mais torres estejam sob alcance do dispositivo, pode-se estimar uma posição mais precisa através de triangulações e ângulos de recepção de sinal [Naboulsi et al. 2016]. Por sua vez, a localização de dispositivos conectados a pontos de acesso à Internet (*Access Points*) é dada pelo identificador do ponto de acesso e o horário no qual o acesso foi iniciado. Assim como nas redes móveis, a precisão da posição informada corresponde ao raio de alcance do ponto de acesso, e múltiplos pontos de acesso podem ser usados para obter uma localização mais precisa.

Existem algumas vantagens no uso de redes sem fio para a coleta de dados geoespaciais. Primeiro, podemos destacar o consumo de energia deste método, dado que as coletas dependem somente da conexão dos dispositivos à rede, atividade fundamental para o seu uso. Além disso, por possuir menor precisão, é menos invasiva que a coleta através de sensores GPS, o que reduz a rejeição à coleta por parte dos indivíduos sensoreados. Por fim, podemos destacar também o número maior de dispositivos capazes de se

conectar às redes sem fio em comparação ao número de dispositivos com sensores GPS. Porém, é válido ressaltar a precisão inferior da coleta através das redes móveis, fator que influencia na utilidade dos dados coletados.

3.3.1.3. Dispositivos *beacons*

Os dispositivos *beacons* são transmissores de curto alcance que permitem monitorar a presença de dispositivos móveis (como *smartphones* e *tags* eletrônicas) através de sinais de proximidade, que podem variar de 5 a 100 m, dependendo da aplicação desejada. Assim, é possível afirmar a posição do dispositivo, com precisão relativa ao raio de alcance do transmissor. Adicionalmente, múltiplos *beacons* podem ser usados para aumentar a precisão da posição reportada.

Os *beacons* são pequenos, portáteis e de alta eficiência energética, devido ao uso de tecnologias de transmissão de baixo consumo, o que permite a sua instalação em diversos locais, sem a necessidade de infraestruturas adicionais. Graças a sua característica de detectar dispositivos com base na proximidade e ao raio de alcance, eles são comumente usados para localização em ambientes internos, como shoppings, supermercados e escritórios, permitindo mapear a presença em andares, salas e corredores. Um número maior de *beacons* significa, portanto, um sensoriamento mais detalhado. A principal desvantagem de seu uso para a coleta de dados geoespaciais está no custo dos dispositivos, que somado à quantidade elevada de dispositivos necessários para o sensoriamento detalhado dos ambientes torna a sua implementação difícil.

3.3.1.4. Endereços postais

Os endereços postais são uma forma detalhada de representar localizações como residências, locais de trabalho e outros locais de interesse. Eles são especificados de acordo com as normas de apresentação de cada país e, no Brasil, por exemplo, são compostos por logradouro (rua ou avenida), número da residência, complemento, nome do bairro, da cidade e do estado e código de endereçamento postal (CEP). Sua precisão é diretamente proporcional à menor unidade (número, rua, bairro, etc) do endereço coletado. Além disso, são mapeáveis para coordenadas em latitude e longitude aproximadas (e vice-versa) através de bibliotecas e *APIs* como *Google Maps Geocoding*[5]. Este tipo de dado geoespacial está presente em uma variedade de sistemas gerenciais e comerciais, especialmente aqueles que envolvem o cadastro de pessoas e a necessidade de serviços postais. Nos últimos anos, eles têm sido usados para mapear os locais de interesses reportados por usuários em *LBSNs*, gerando como resultados registros de *check-in*.

Comparados às outras fontes de dados, os endereços postais têm como vantagem a sua facilidade de leitura e interpretação e o fato de não precisarem de dispositivos sensores específicos para a coleta da localização. Porém, por outro lado, os dados de endereços postais são úteis somente para reportar localizações fixas relacionadas ao usuário (como residência e local de trabalho), não servindo para reportar posições específicas sem en-

[5]https://cloud.google.com/maps-platform/places

dereço definido (como um ponto remoto na superfície da Terra) e posições instantâneas entre dois locais definidos (e.g., os pontos que compõem a trajetória de um usuário).

3.3.2. Qualidade da coleta

Além das tecnologias utilizadas e das fontes de dados geoespaciais, outras questões devem ser levadas em consideração durante a definição do escopo da coleta dos dados. A seguir, definimos essas questões e discutimos como as mesmas podem afetar, não somente a coleta dos dados, mas também o processamento e os resultados das análises.

3.3.2.1. Acurácia e Precisão

A acurácia, no âmbito da coleta de dados geoespaciais, diz respeito à proximidade da medida de localização coletada de um valor aceitável para a posição real de uma entidade. Portanto, quanto maior a acurácia, mais próximo a medida está do valor correto. A acurácia dos tipos de dados varia de alguns milímetros (sensores GPS de alta capacidade) a quilômetros (sensores de redes móveis em áreas remotas), portanto é fundamental compreender o conceito de acurácia para que se possa escolher o tipo de dado geoespacial que mais se ajusta à análise a ser feita.

Por sua vez, a precisão representa a variação das amostras coletadas e, quanto maior o seu valor, mais centrados estão as amostras em torno de um ponto. Para obter medidas mais precisas, sensores mais potentes podem ser aplicados, resultando porém em um maior consumo de energia e um custo maior da infraestrutura de coleta.

A Figura 3.4 apresenta o significado de acurácia e precisão (e suas variações) quando tratamos de dados geoespaciais. Nota-se que obter uma alta precisão não necessariamente implica em resultados acurados. Adicionalmente, uma alta acurácia pode produzir resultados nem sempre confiáveis se a precisão for baixa.

Na presença de dados com baixa acurácia ou de sensores imprecisos, duas abordagens podem ser aplicadas. A primeira delas é fazer o uso de sensores mais potentes – quando possível. A segunda é através de técnicas que podem ser aplicadas para corrigir as medidas sensoreadas. Isto pode ocorrer durante a coleta, como no caso do uso de sinais adicionais no sistema GPS e nas redes sem fio, e também durante o pré-processamento dos dados. Quanto a este último, existem abordagens para a correção de dados com baixa acurácia [Newson e Krumm 2009, Hoteit et al. 2016].

3.3.2.2. Privacidade

Dados geoespaciais são utilizados para reportar a localização de várias entidades, com destaque para aquelas capazes de se locomover, como humanos e veículos. Assim, ao fornecer sua posição, um indivíduo permite o acesso a uma informação real e de grande validade. Por um lado, são inúmeros os benefícios provindos das análises de dados de mobilidade. Por outro lado, essas informações podem ser utilizadas por agentes mal-intencionados para diversos ataques à privacidade dos usuários sensoreados, o que pode desmotivá-los a compartilhar seus dados. O compromisso entre a utilidade dos dados

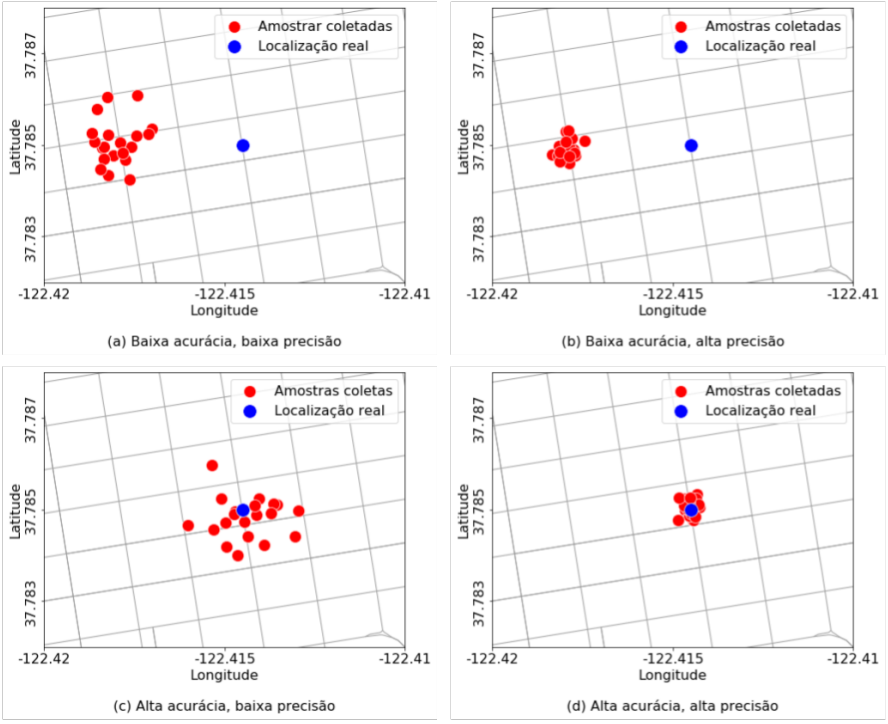


Figura 3.4. Variações de acurácia e precisão de dados geoespaciais coletados e os seus significados

geoespaciais e o risco à privacidade dos usuários deve ser considerado ao realizar a coleta dos dados.

Para que os usuários possam compartilhar suas informações de localização, é preciso garantir a sua privacidade com o uso de técnicas que reduzem os detalhes dos dados compartilhados ou que dificultam o seu acesso. Podemos destacar duas técnicas: a anonimização e a obsfucação dos dados. Quanto à primeira, substiui-se os identificadores dos usuários por pseudônimos gerados aleatoriamente, o que pode ocorrer tanto durante o processamento dos dados quanto diretamente nos dispositivos coletores [Krumm 2009]. Porém, mesmo com a anonimização, ataques podem reidentificar usuários a partir de seus locais de interesse [Maouche et al. 2017]. O papel da obfuscação é impedir que esse tipo de ataque seja possível, realizando pequenas distorções nas posições reportadas, sem alterar a qualidade dos dados [Duckham e Kulik 2005a, Duckham e Kulik 2005b].

3.3.2.3. Frequência

O intervalo entre duas coletas consecutivas de dados também é um importante aspecto a ser discutido. Chamamos esse intervalo de frequência de amostragem, de forma que dados com alta frequência são aqueles com um menor intervalo de tempo entre duas amostras, e dados com baixa frequência são aqueles com um maior intervalo de tempo entre duas amostras. Portanto, durante um mesmo intervalo de tempo, frequências maiores produzem uma quantidade maior de dados. Quanto mais dados, mais detalhes existem e maior é a sua utilidade, podendo ser empregados para estudar trajetórias detalhadas de desloca-

mento, por exemplo. Por outro lado, coletar dados com alta frequência leva a um consumo maior de recursos como energia e armazenamento, que são limitados em dispositivos portáteis. Ao implementar um processo de coleta de dados geoespaciais, é preciso especificar a frequência da coleta de modo a garantir a cobertura das atividades ou comportamentos de interesse da análise.

Enquanto alguns tipos de dados permitem a definição da frequência da coleta (e.g., sistemas GPS), outros dependem da interação do usuário sensoreado com o sistema, e portanto sua frequência de coleta não pode ser controlada diretamente. Com isso, surgem lacunas espaciais e temporais, que são intervalos extensos de espaço e tempo, respectivamente, onde não se têm conhecimento das atividades do usuário. Essas lacunas podem ser preenchidas utilizando técnicas como interpolação e extrapolação [Hoteit et al. 2016], algoritmos baseados em dados históricos [Silva et al. 2015] e aprendizado de máquina [Chen et al. 2017]. O enriquecimento dos dados geoespaciais transforma dados esparsos em dados densos de forma artificial, sem que mudanças nos métodos e ferramentas de coleta sejam necessárias.

3.3.2.4. Escala

Por fim, discutimos a escala dos dados geoespaciais. Para representar o ambiente simulado e produzir resultados significativos, é necessário que o conjunto de entidades presentes nos dados seja suficiente. A escala pode se referir ao número de usuários distintos monitorados, ao número de intervalos de tempo (horas, dias ou semanas) e às dimensões da região física monitorada. As escalas destas dimensões devem compreender um ambiente de simulação no qual os resultados não sofram viés devido a limitações de usuários (o conjunto de indivíduos não representa a população), de tempo (o período coberto não engloba todas as situações esperadas pelo experimento) e de espaço (as dimensões da região coletada não se assemelham àquelas do ambiente real).

Quando os dados geoespaciais coletados não são suficientes para produzir resultados, algumas abordagens são usadas para aumentar o volume dos dados. A fusão de dados [Rettore et al. 2020] é uma técnica que permite a junção de dois ou mais conjuntos de dados, gerando como saída um único conjunto contendo todos os dados. Em conjuntos de dados com estrutura semelhante, sua implementação é simples. Porém em casos onde a interseção entre as partes é pequena ou inexistente, a sua aplicação demanda processamento elevado. Já a geração de dados sintéticos faz uso de métodos estatísticos [Kosta et al. 2012] e de aprendizado de máquina para gerar dados sintéticos realísticos, i.e., que se aproximam do comportamento real. Apesar de demandar uma modelagem precisa dos dados reais, essa técnica tem como benefício a capacidade de gerar dados sintéticos sob demanda e em alta escala.

3.4. Armazenamento

Esta seção irá abordar os conceitos, técnicas e as ferramentas existentes para o armazenamento de dados geoespaciais, que, como outros tipos de dados, devem ser armazenados após a coleta para serem utilizados posteriormente. É importante discutir sobre esse assunto pois, muitas vezes, as formas de armazenamento tradicionais, como bancos de

dados relacionais, não são as mais apropriadas para este tipo de dado. Isso ocorre devido às diferentes formas de representação existentes para dados geoespaciais que não são compatíveis com o armazenamento de dados tabulares. Além disso, deve-se levar em consideração o propósito da manipulação dos dados, com informações sobre a frequência de consultas e inserções, a necessidade de realizar filtros geoespaciais de forma eficiente e o volume de dados que será armazenado.

3.4.1. Estrutura de Componentes Espaciais

O mundo real é muito complexo para ser representado completamente por uma estrutura de dados, sendo necessária a escolha das feições (e.g., águas, estradas, árvores, etc.) relevantes para cada caso. Para a representação digital de dados geoespaciais, existem duas estruturas primárias: a vetorial e a matricial (também chamada de *raster*). A estrutura vetorial baseia-se na utilização de pontos, linhas e polígonos para definir a localização e os limites de um objeto. Por sua vez, a estrutura matricial utiliza uma grade regular de células para definir os objetos. Cada estrutura possui suas vantagens e desvantagens na modelagem de dados. Ademais, pode-se combinar as duas abordagens em um único projeto visando extrair as vantagens de ambas. A Tabela 3.2 apresenta uma comparação entre os dois tipos.

Tabela 3.2. Comparação entre os Modelos Vetorial e Matricial (adaptado de [Marino 2012])

Característica	Vetorial	Matricial
Estrutura de Dados	Geralmente Complexo	Geralmente Simples
Requisito de Armazenagem	Pequena, para maior parte dos dados	Grande para a maioria dos dados sem compressão
Conversão de Sistema de Coordenadas	Simples	Pode ser lenta, devido ao volume, e requerer reamostragem
Precisão Posicional	Limitado pela quantidade posicional de levantamento	Degraus contornando células; depende da resolução adotada
Acessibilidade	Frequentemente complexo	Fácil para modificar através do uso de programas
Visualização e Saída	Parecido com mapas, com curvas contínuas; pobre para imagens	Bom para imagens, mas para feições discretas, pode mostrar efeito escada
Relações espaciais entre objetos	Relacionamentos topológicos entre objetos disponíveis	Relacionamentos espaciais devem ser inferidos
Análise e Modelagem	Álgebra de mapas é limitada	Superposição e modelagem mais fáceis

As escolhas na construção de uma estrutura impactam diretamente nos detalhes que elas são capazes de capturar. Na Figura 3.5 são apresentadas as transformações de uma representação do mundo real para as estruturas vetorial e matricial. Nela, é possível

ver as características de modelagem de cada uma das estruturas. É possível notar que a estrutura vetorial representa as entidades existentes considerando suas dimensões e seu formato. Para isso, deve-se escolher quais as formas geométricas que serão utilizadas na representação. Por sua vez, a estrutura matricial reduz todas as feições contidas em uma única célula a uma identificação básica de acordo com um critério de codificação. Na Figura 3.5 utilizou-se o critério de área dominante, no qual o rótulo corresponde à feição que ocupa a maior parte da célula, podendo haver um identificador no caso de combinações de feições. Podemos destacar também o método de centro da célula, no qual o rótulo é proveniente da feição presente no centro da célula. Por fim, o método de cobertura percentual constrói uma matriz para cada feição, e o rótulo da célula é definido pela porcentagem daquele elemento.

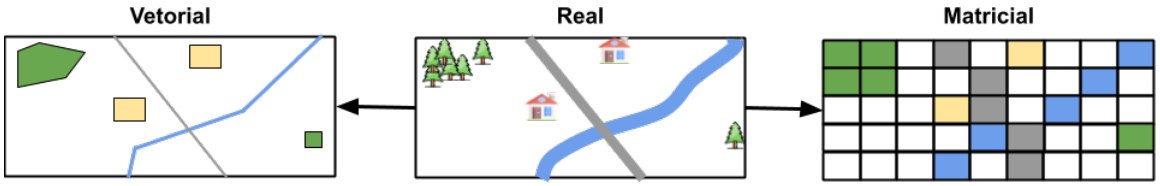


Figura 3.5. Estrutura Vetorial vs. Matricial (baseada na Figura 5.5 de [Lisboa Filho e Iochpe 2001])

Levando em consideração as características que foram citadas anteriormente, podemos perceber que uma estrutura não é melhor que a outra. A estrutura matricial, além de ser mais simples para armazenar, possui a facilidade na aplicação de operações entre camadas. Outros motivos para escolher esta estrutura seria para o armazenamento, visualização e manipulação de imagens digitais, como fotografias aéreas e imagens de satélites, por exemplo. Já a estrutura vetorial é, de um modo geral, mais parecida com os mapas e com isso, na maioria das vezes, possui os dados mais acurados. Por estes motivos, a estrutura vetorial permite uma visualização mais harmoniosa e um cálculo mais eficiente de operações topológicas. Por fim, a estrutura vetorial armazena somente os elementos essenciais enquanto a matricial codifica cada célula, o que às vezes é desnecessário. Existem várias outras estruturas, como a Rede Triangular Irregular (*Triangulated Irregular Network* – TIN) [Longley et al. 2005], por exemplo. Apesar de ser mais complexa que as outras, a estrutura TIN é melhor para representar superfícies, como elevações, por meio da junção de triângulos.

3.4.2. Compactação dos Dados

Como pode ser notado, os conjuntos de dados geoespaciais tendem a representar uma grande quantidade de informação e, portanto, necessitam de capacidades de armazenamento consideráveis. Assim como para os conjuntos de dados tradicionais, algoritmos de compactação de dados podem ser aplicados nos conjuntos de dados geoespaciais, resultando em um armazenamento mais eficiente. Existem vários algoritmos para comprimir os mais diversos tipos de arquivos, podendo ser classificados em compactação com perdas e compactação sem perdas. Por um lado, a compactação com perdas obtém altos níveis de compactação, porém ao custo de perdas na qualidade dos dados. Por outro lado, a compactação sem perdas não afeta a qualidade dos dados, mas tem menor eficiência.

Embora para certas aplicações a perda de informação é aceitável, quando se trata de dados geoespaciais ela pode impactar gravemente na qualidade da amostragem. Assim, não se recomenda utilizar qualquer tipo de compactação que gere perdas quando se deseja processar ou analisar os dados. Por este motivo, é mais comum aplicar algoritmos de compactação em dados matriciais discretos [Longley et al. 2005].

Um método comum para compactação de dados matriciais é o *Run-length code*. Essa técnica de compressão se baseia em codificar a sequência de células a fim de otimizar o espaço quando há grandes repetições de células com mesmo valor. Essa codificação é representada por dois números, sendo que o primeiro indica a quantidade de células com a mesma identificação e o segundo a identificação comum em si. A Figura 3.6 ilustra um exemplo do uso do *Run-lenght code*, onde é possível ver a redução que pode ser alcançada com essa técnica.

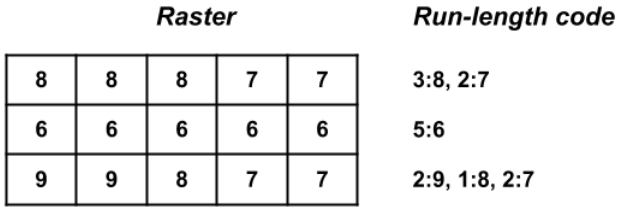


Figura 3.6. Codificação *Run-length code* de um *raster*

Outra codificação bastante conhecida é uma versão bidimensional da *Run-length code*, a *Quad tree* [Finkel e Bentley 1974]. Neste método, áreas que possuem os mesmos valores são representadas com um único identificador. Para isso, a matriz é dividida recursivamente em blocos quadrados, de tamanho sempre crescente, até que não seja mais possível dividi-la, resultando em quadrados em que todos os identificadores que o compõem são iguais. A Figura 3.7 mostra como seria a aplicação desta codificação em um dado matricial.

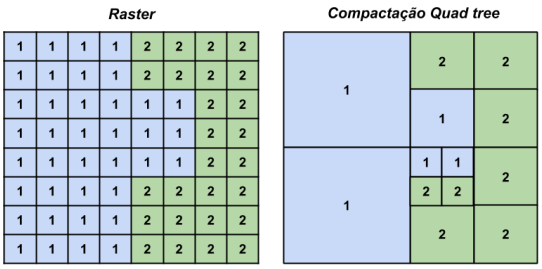


Figura 3.7. Codificação *Quad tree* de um *raster*

Os métodos apresentados reduzem a redundância para um armazenamento mais eficiente. Contudo, existes métodos como *wavelet* que permite perdas, já que parte dos dados são descartados durante a compactação. Apesar deste tipo de técnica resultar em altas taxas de compressão, dada a perda de informação, sua utilização é limitada a projetos que não utilizam dados brutos para análise. Também por causa da perda, normalmente, não se aplica este tipo de compressão em estruturas vetoriais. Isso porque, esta estrutura

já trabalha com os dados na forma reduzida, descartando elementos irrelevantes já no processo de coleta [Longley et al. 2005]. Entretanto, é possível diminuir a redundância ao utilizar a estrutura vetorial quando a área mapeada é pequena. Isso pode ser feito armazenando somente um ponto no mapa com as coordenadas completas, enquanto o restante, ao invés de possuírem coordenadas absolutas, seriam mapeados utilizando como referência este ponto base [Claudia Dolci 2010].

3.4.3. Bancos de Dados

Um dos componentes mais importantes para Sistemas de Informações Geográficas (SIGs) são os Sistemas de Gerenciamento de Banco de Dados Geográficos (SGBDGs). Além de possuírem as funcionalidades de armazenamento encontradas em sistemas de gerenciamento de banco de dados tradicionais, os SGBDGs aceitam diferentes sistemas de referência geoespacial, provendo funções próprias para realização de busca e manipulação desse tipo de dado. Adicionalmente, são capazes de indexar dados geoespaciais tanto em formato de coordenadas quanto em formato de polígonos, aprimorando a eficiência geral do sistema. No entanto, sem o uso de indexação, tanto a busca por localizações quanto o filtro de regiões serão muito ineficientes, principalmente na presença de grandes volumes de dados.

A Tabela 3.3 apresenta três SGBDGs que são comumente usados para o armazenamento de dados geoespaciais: MySQL [mys 2020], Postgree com a extensão PostGis [Developers 2020] e Oracle Spatial [Oracle 2020]. Podemos observar que todos utilizam o mesmo padrão de tipo espacial, o *Simple Feature Specification - Structured Query Language* (SFS-SQL), que descreve um modelo comum de armazenamento e acesso de geometrias (pontos, linhas e polígonos). O SFS-SQL é um padrão definido pela *Open Geospatial Consortium* (OGC) [ogc 2020] e que, além de descrever as geometrias utilizadas pelos SIGs, apresenta definições de operações como *touches, equals, overlaps, disjoints, intersects*, dentre outras. Quanto à transformação de sistemas de coordenadas, somente o MySQL não a fornece, sendo necessário realizar as conversões entre sistemas antes do armazenamento dos dados. Quanto à indexação dos dados geoespaciais, os três sistemas analisados fazem uso da indexação *R-Tree*. Por fim, somente o Oracle Spatial possui custos de utilização, o que é compensado pela robustez do sistema, mas mesmo assim [Shukla et al. 2016] mostra que para a maioria das operações o PostGIS é mais otimizado que o Oracle Spatial. Portanto, para escolha do SGBDG a ser utilizado, deve-se considerar os objetivos do projeto e suas características.

Tabela 3.3. Comparação entre os SGBDs com extensão SIG. (Adaptado de [Casanova et al. 2005])

Característica	MySQL	PostGIS	Oracle Spatial
Tipos Espaciais	SFS-SQL	SFS-SQL	SFS-SQL
Transformação de sistemas de coordenadas	Não	Sim	Sim
Indexação	R-Tree	R-Tree sobre Gist	R-Tree e QuadTree
Custo	Gratuito	Gratuito	Pago

3.4.4. Indexação

Índices são estruturas de dados usadas para aumentar o desempenho de consultas em sistemas de banco de dados, permitindo a localização de dados de forma mais eficiente do que buscas lineares. A indexação (isto é, a criação de índices) em bancos de dados espaciais é útil não somente para a recuperação eficiente de dados, mas também para diversas operações espaciais. Podemos citar a identificação dos k pontos mais próximos (*k-Nearest Neighbors*), a geocodificação (obtenção das coordenadas a partir de informações como endereço) e a geocodificação reversa (obtenção de informações a partir de uma coordenada). A seguir, utilizaremos o problema da geocodificação reversa como estudo de caso para demonstrar as estruturas de indexação de dados geoespaciais presentes na literatura.

O problema de múltiplos pontos em múltiplos polígonos é importante no processo de geocodificação reversa, ilustrado na Figura 3.8. Podemos descrevê-lo da seguinte maneira: Dado um conjunto $N = \{n_0, n_1, \cdots, n_I\}$ de polígonos e um conjunto $M = \{m_0, m_1, \cdots, m_J\}$ de pontos em um plano, deseja-se saber em qual polígono n_i cada um dos m_j pontos está contido. Para um conjunto pequeno de dados, a solução por força bruta é suficiente: para cada polígono, verificamos se cada um dos pontos está contido nele. Como um ponto só pode estar em uma região por vez (assumindo polígonos sem interseção), os pontos que já foram encontrados podem ser retirados da busca.

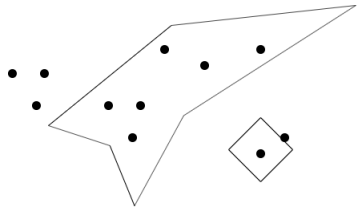


Figura 3.8. Problema do ponto em polígono: Dado um conjunto de pontos e polígonos, deseja-se saber em qual polígono cada ponto está contido.

Algoritmos clássicos de geometria computacional, como o chamado *raycasting* [Shimrat 1962], resolvem esse problema quando é necessário determinar se somente um ponto está contido em um polígono. Dessa forma, sucessivas aplicações do algoritmo resultarão na resolução do problema completo. Entretanto, essa abordagem possui custo computacional alto, o que inviabiliza sua utilização para grandes volumes de dados.

A estrutura de dados mais aplicada na resolução desse problema é a chamada R-Tree [Guttman 1984], que permite a indexação de geometrias através de uma estrutura de árvore balanceada. Essa estratégia tem como vantagem a busca com complexidade logarítmica de tempo ($O(\log_{|M|} |N|)$, o que justifica o seu uso nas implementações de diversos bancos de dados e ferramentas de análises geoespaciais, como PostGIS, Oracle e GeoPandas. Por outro lado, é preciso levar em consideração o alto custo de tempo e espaço para a construção da árvore, fator que pode limitar o seu uso. Além da geocodificação reversa, a R-Tree também é utilizada para resolver outros problemas, como o *k-Nearest Neighbors*. A R-Tree acelera a geocodificação reversa indexando as caixas delimitadoras de cada polígono. Com a indexação das caixas delimitadoras é possível verificar em qual polígono algum ponto está contido. Isso se deve ao fato de um ponto não poder estar contido em um polígono, se não estiver contido em sua caixa delimitadora.

Além da R-Tree, podemos utilizar os sistemas de grades como o chamado *Geohash* [Morton 1966] e o H3 [Uber 2015] para esse problema. Os sistemas de grade permitem analisar grandes conjuntos de dados espaciais através da divisão de áreas maiores em células unicamente identificáveis. O H3 provê um índice espacial hierárquico baseado em hexágonos, que permite agrupar pontos em hexágonos de diversos tamanhos, conforme a necessidade de precisão da análise. O nível do índice H3 determina a área dos hexágonos e sua escolha é essencial para uma melhor precisão na indexação. Por um lado, hexágonos muito grandes agruparão pontos mais distantes em uma mesma célula. Por outro lado, hexágonos muito pequenos resultarão em um número muito grande de índices, afetando o desempenho. Para isso, os desenvolvedores dessa tecnologia disponibilizam uma tabela com a média da área de um hexágono de acordo com a precisão necessária[6].

O *GeoHash*, ao invés de utilizar hexágonos, utiliza retângulos. Porém, o princípio para a indexação é o mesmo do H3, com a desvantagem de existirem oito vizinhos para cada retângulo em comparação com seis no caso do hexágono. Além disso, no caso dos hexágonos, os pontos centrais de todos os vizinhos são equidistantes do hexágono central.

Para resolver o problema de geocodificação reversa, é preciso criar uma tabela de pesquisa com uma coluna contendo índices H3 ou GeoHash, em uma precisão adequada para os tamanhos do polígonos, e outra coluna contendo um identificador único para cada polígono. Esse identificador pode ser o número de uma cidade ou setor censitário fornecido pelo IBGE, por exemplo. Com essa tabela, um ponto pode ser mapeado em um polígono ao se consultar na tabela o índice H3 ou GeoHash desse ponto.

Semelhante ao H3, o Appel [Coimbra et al. 2019] é um sistema que apresenta uma estratégia de geocodificação reversa baseada em uma estrutura de dados hierárquica de polígonos. Nele, cada nível da árvore representa um tipo de subdivisão geográfica do Brasil (e.g., estados, mesorregiões e microrregiões). As regiões em um mesmo nível são disjuntas, isto é, não possuem interseção entre si e, portanto, os pontos pertencentes a uma região estariam contidos em somente um polígono. Cada nó, com exceção dos nós folha, contém uma lista de polígonos que compõem a região no nível superior. Esta lista, por sua vez, é ordenada de acordo com a população residente em cada polígono, com o intuito de acelerar a busca linear dentro de um nó.

Finalmente, o *Elastic Search* é um sistema de pesquisa e análise de dados que tem atraído bastante atenção recentemente. Além de outras aplicações, o *Elastic Search* permite a análise de dados geoespaciais, como por exemplo as operações de junção espacial. Nesse caso, ele cria os índices de formas geométricas decompondo-as em uma malha de triângulos, e indexando cada triângulo como um ponto de sete dimensões em uma árvore KBD (*K-Dimensional B-Tree*). Essa abordagem de tesselação permite uma precisão elevada, enquanto o desempenho dependerá do número de vértices dos polígonos.

A Tabela 3.4 compara o tempo necessário para localizar uma quantidade de pontos em polígonos de setores censitários brasileiros. É importante notar que o PostGIS utiliza a R-tree otimizada para atuar em memória secundária, enquanto que o GeoPandas utiliza a R-tree em memória. Os códigos H3 e GeoHash foram carregados em um servidor Redis, que os mantém em memória. Já a Tabela 3.5 apresenta outras métricas de comparação,

[6]https://h3geo.org/#/documentation/core-library/resolution-table

como a memória gasta para armazenar o modelo de consulta e o tempo necessário para gerar os modelos.

Tabela 3.4. Comparação de tempo de consulta para a geocodificação reversa nas ferramentas de indexação. Adaptada de [Coimbra et al. 2019]

Pontos	APPEL (s)	*PostGIS* (s)	*GeoPandas* (s)	H3	GeoHash
10^4	$0,82 \pm 0,01$	$4,30 \pm 0,24$	$4,52 \pm 0,02$	$0,14 \pm 0,01$	$0,23 \pm 0,01$
10^5	$2,80 \pm 0,01$	$44,21 \pm 0,92$	$7,59 \pm 0,06$	$1,80 \pm 0,04$	$2,05 \pm 0,11$
10^6	$20,92 \pm 0,02$	$439,77 \pm 10,73$	$72,47 \pm 0,57$	$18,140 \pm 0,59$	$21,29 \pm 0,69$

Tabela 3.5. Comparação nos tempos e tamanhos dos modelos gerados para as consultas.

Métrica	APPEL	*PostGIS*	*GeoPandas*)	H3	GeoHash
Geração do modelo (min)	156,50	0,01	1,73	23,59	80,08
Tamanho do Modelo (MB)	7,34	19,56	2,34	5,10	5,30

3.5. Preparação e Extração de Conhecimento

Esta seção apresenta o núcleo do processo de análise de dados geoespaciais, que compreende as etapas de preparação dos dados e extração de conhecimento.

3.5.1. Preparação

A preparação de dados geoespaciais envolve cinco etapas: formatação, amostragem, limpeza, filtragem e agregação. Apesar de todas possuírem papel essencial na preparação dos dados para a análise, a necessidade de aplicação de cada uma é definida pelas condições iniciais dos dados de entrada e pelas características definidas para o escopo do trabalho. Ademais, múltiplas iterações de preparação dos dados podem acontecer com o intuito de refinar ou validar os resultados obtidos.

3.5.1.1. Formatação

Ao trabalhar com dados geoespaciais, uma vez que estes sejam recuperados, é essencial atentarmos à sua formatação, já que dados geoespaciais podem ser representados de diversas formas. No que compete à representação de coordenadas geográficas (latitude e longitude), existem três formatos básicos em que estes podem ser encontrados: a utilização de graus, segundos e minutos; graus, minutos e decimais do minuto; e graus e decimais do grau. A escolha de qual formato utilizar dependerá do propósito da aplicação, dado que um número maior de casas decimais permite representar localizações com maior precisão.

A formatação que utiliza graus, minutos e segundos (DDD° MM' SS") é normalmente utilizada para representação de coordenadas em mapas, porém é mais difícil de ser trabalhada em sistemas computacionais. Já a utilização de graus, minutos e decimais do minuto (DDD° MM.MMM') se dá principalmente em equipamentos de navegação eletrônica. Apesar da existência dos formatos anteriores, boa parte das ferramentas e bibliotecas

mais frequentemente utilizadas em sistemas computacionais trabalha com o terceiro formato, que trata as coordenadas como graus e decimais do grau (DDD.DDDDD°). Entre estas ferramentas estão SGBDGs, softwares SIG, além de diversas fontes de dados oficiais disponibilizadas em formato de arquivo. Isso acontece porque este formato já está pronto para utilização, não necessitando de qualquer conversão. Vale lembrar também que cada fonte de dados pode utilizar um padrão diferente no que compete à ordem na qual latitude e longitude são apresentadas, sendo necessária atenção a este fator para a representação correta dos pontos.

Outra representação possível é a de conjuntos de coordenadas, representando polígonos ou linhas. O formato adotado para este fim dependerá da ferramenta utilizada ou da fonte dos dados. Por exemplo, um dado obtido de uma plataforma como *OpenStreetMaps* pode representar um polígono simplesmente como uma sequência de coordenadas, enquanto que um polígono representado pela biblioteca *Shapely* utiliza o formato *Well-Known Text* (WKT) para representar formas geométricas como uma sequência de coordenadas. Para isso, o WKT define um conjunto de palavras para representar objetos distintos, tais como *'POINT()'*, *'LINESTRING()'* e *'POLYGON()'*.

É válido ressaltar que algumas fontes de dados os disponibilizam em forma de *Shapefiles* ou ainda como imagens. Alguns sistemas SIG podem trabalhar com ambas as abordagens, sendo que a leitura de imagens normalmente se dá para análise de dados matriciais. Já bibliotecas como *Geopandas* e *Fiona* trabalham somente com formatos textuais, como *Shapefiles*, planilhas, *JSON* e WKT.

Além de se certificar que os dados possuam os formatos corretos, é importante também verificar se todos os dados utilizam o mesmo tipo de projeção ou se estão no mesmo *datum*, pois dados que utilizam de projeções diferentes podem levar a análises espaciais incorretas. Se esse for o caso, a biblioteca *Geopandas* e outros sistemas SIG, como o QGIS, por exemplo, possuem meios para que a correção possa ser feita.

3.5.1.2. Limpeza

Além de localizações indesejadas, a utilização de dados comprovadamente em área urbana pode trazer alguns desafios, tais como a imprecisão das coordenadas geográficas. Isso acontece devido ao grande número de edifícios que obstruem a visão dos satélites, fazendo com que estes não consigam atribuir uma localização com a exatidão requerida. Esse fenômeno é denominado *urban canyons*, e pode ocorrer não somente devido a edifícios como também a túneis e trincheiras [Johnson e Watson 1984]. Por este motivo, é importante que se analise o impacto das localizações afetadas nos dados coletados e, caso seja necessário, que seja feita a exclusão desses dados.

Uma das consequências do uso de dados com baixa acurácia de localização é a impossibilidade de se utilizar algoritmos que trabalham com a distância entre pontos ou a densidade de regiões. Isso acontece porque a localização geográfica dos pontos será de vital importância para que os resultados sejam computados da forma correta. Ou seja, se um ponto está referenciado a uma região que não corresponde à sua posição no mundo real, o resultado produzido pelo algoritmo não retratará a realidade. Podemos observar o problema com maior clareza ao analisarmos um sistema para mobilidade urbana. Se um

passageiro solicita um transporte de sua casa até o trabalho mas o local de origem (ou destino) está georreferenciado de maneira errônea, o sistema tenderá a cobrar um preço diferente daquele que seria o correto, causando prejuízo para uma das partes.

É válido ressaltar que, mesmo que não se refira a dados geoespaciais, é necessária a atenção durante a limpeza às outras dimensões dos dados coletados. Essas, apesar de não serem afetadas pelas questões presentes nos dados geoespaciais (como precisão dos sensores), podem conter irregularidades como valores nulos ou fora do intervalo esperado (*outliers*). Sendo assim, a sua limpeza é essencial para garantir que as análises sejam feitas somente com dados válidos.

3.5.1.3. Filtragem

Enquanto a etapa de limpeza dos dados tem como foco a remoção de informações sensoreadas erroneamente, a etapa de filtragem visa selecionar, dado um conjunto de dados limpos, um subconjunto que atenda a regras especificadas para a análise. Neste capítulo nos referimos às regras aplicáveis às dimensões geoespaciais dos dados, mas é válido ressaltar que a filtragem pode contemplar as diversas dimensões presentes no conjunto. Assim, de acordo com a aplicação dos dados, podemos considerar dispensável o uso de dados localizados em regiões ermas (e.g., oceanos, desertos, florestas, entre outras), ou ainda pontos que estejam fora de uma área específica (fora dos limites de uma cidade ou estado, por exemplo). Aqui, ressaltamos que muitas vezes a representação geométrica do dado vem na forma de um ponto, ou seja, um registro com latitude e longitude. Para realizar a filtragem de registros somente em locais de interesse, podemos considerar uma representação geométrica da área de interesse. Se desejamos coletar somente os dados de usuários em uma avenida, por exemplo, podemos utilizar a representação geométrica de tal avenida, que poderá ser uma linha ou um polígono, para filtrar os pontos. Já para um estado, por sua vez, podemos utilizar o polígono que representa suas delimitações.

Para isso, várias operações geográficas podem ser aplicadas para testar o relacionamento entre geometrias, sendo cada uma mais indicada para um tipo de aplicação. Tais operações estão disponíveis em SIG e em bancos de dados com extensão para dados geográficos, além de bibliotecas como *Shapely* e *GeoPandas*. A seguir, detalhamos as principais operações, sendo que cada uma retorna como saída um valor *booleano* (verdadeiro ou falso) para o relacionamento entre as representações geométricas [Longley et al. 2005]. Além disso, cada tipo de operação pode ser válida ou não para a combinação de elementos geométricos analisados. A Figura 3.9 ilustra casos onde cada operação irá retornar VERDADEIRO ou FALSO, sendo que cada uma recebe como parâmetro duas geometrias (A e B). Para melhor visualização, as representações em preto dizem respeito à geometria A, e as demais à geometria B. Além disso, os resultados são obtidos através da expressão A.OPERACAO(B), onde OPERACAO representa as operações abaixo.

- ***CONTAINS:*** Verifica se uma representação contém completamente a outra. Inválida para a combinação ponto-linha, pois uma linha não pode estar completamente contida dentro de um ponto; porém a combinação inversa é válida.
- ***CROSSES:*** Analisa se as representações se sobrepõem em algum lugar, ou seja,

se as geometrias possuem pontos interiores em comum, mas não todos (uma não está contida na outra). Vale ressaltar que esta operação pode ser usada para representações com quantidade de dimensões diferentes, por exemplo uma linha e um polígono.

- ***DISJOINT:*** Verifica se as representações utilizadas são disjuntas, ou seja, não compartilham nenhum ponto em comum.
- ***EQUALS:*** Verifica se as duas geometrias são iguais.
- ***INTERSECTS:*** Analisa se as geometrias se interceptam em algum ponto, ou seja, compartilham qualquer porção de espaço. Retorna FALSO se as geometrias forem disjuntas.
- ***OVERLAPS:*** Analisa se representações de mesma dimensão se sobrepõem, mas uma não está contida na outra.
- ***RELATE:*** Verifica de forma mais geral se duas representações se relacionam através de interseções nos limites, interiores ou exteriores desta, mas não são disjuntas. Esta operação é útil para verificar de uma só vez se há interseção ou se as geometrias se cruzam ou se tocam, por exemplo.
- ***TOUCHES:***Analisa se há interseção entre os limites das geometrias, mas seus interiores não se intersectam.
- ***WITHIN:*** Verifica se uma geometria está dentro da outra. Representa a relação inversa de *CONTAINS*.

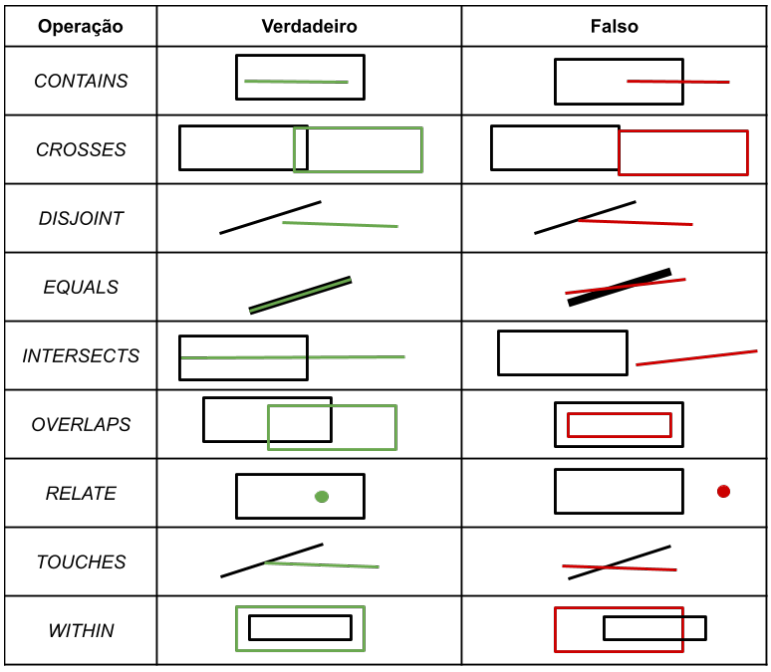


Figura 3.9. Exemplos envolvendo operações que verificam se existe uma relação entre duas geometrias.

Existem ainda outras operações que não analisam somente a relação entre duas geometrias, retornando VERDADEIRO ou FALSO, mas realizam operações espaciais, retornando valores ou novas geometrias como saída, como pode ser visto na Figura 3.10. Tais operações são:

- ***BUFFER*:** Dada uma distância especificada pelo usuário, a operação irá gerar e retornar uma nova geometria resultante da adição de uma silhueta à geometria original.
- ***CONVEXHULL*:** Retorna o envoltório convexo da geometria especificada.
- ***DIFFERENCE*:** Retorna uma geometria que contém todos os pontos que estão na representação de base mas não na geometria de comparação.
- ***DISTANCE*:** Retorna a menor distância possível entre duas geometrias.
- ***INTERSECTION*:** Retorna a geometria que pode ser observada em ambas as representações utilizadas.
- ***SYMDIFFERENCE*:** Retorna a geometria que contém todas aquelas que não se intersectam nas representações utilizadas.
- ***UNION*:** Retorna a geometria obtida com a união de todas aquelas presentes nas duas representações

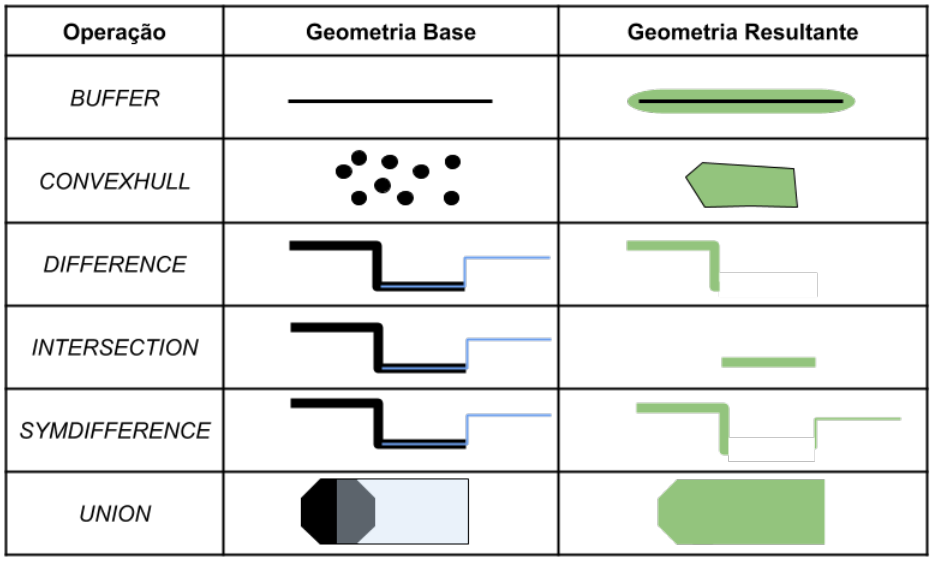


Figura 3.10. Exemplos do resultado obtido quando aplicamos algumas operações à geometrias.

As operações explicitadas acima podem ser encontradas tanto em SGBDs habilitados para uso de dados geoespaciais, quanto em bibliotecas específicas para este fim, como *GeoPandas* e *Shapely*.

3.5.1.4. Amostragem

Uma vez que os dados estão formatados, limpos e filtrados, podemos considerar a utilização de amostras dos dados. Muitas vezes a amostragem dos dados é a primeira etapa a ser concluída, facilitando a limpeza e filtragem dos dados, já que seria necessário lidar com uma quantidade reduzida de dados. Porém, ao realizarmos a amostragem antes destes passos, corremos o risco de os dados da amostra não serem representativos.

Quando analisamos uma grande quantidade de dados, nem sempre possuímos capacidade computacional ou interesse em todos os dados. Muitas vezes, ter uma pequena amostra é suficiente para entendermos e gerarmos conhecimento sobre todos os dados. Entretanto, nem sempre a amostra coletada é representativa, ou seja, possui características semelhantes às da população. Uma amostra ruim pode apresentar viés, o que tornará o conhecimento obtido através dela praticamente descartável. Isto acontece porque não poderemos saber ao certo se a informação obtida é válida para todos os dados ou não.

Em uma amostra válida, os dados são escolhidos aleatoriamente, garantindo que nenhum dado enviesado seja escolhido de forma proposital. Para isso, devemos garantir algumas características. Primeiramente precisamos verificar se a distribuição dos seus dados é originária da distribuição da população, utilizando algum método estatístico, como a estatística de *Kolmogorov-Smirnov* [Massey Jr 1951], por exemplo. Vale lembrar que a distribuição abordada aqui diz respeito à distribuição de todas as informações presentes no conjunto de dados, e não somente dos dados geoespaciais especificamente. Outra opção é recolher diversas amostras e calcular a média de cada uma, pois por definição, a média das amostras irá se acumular em torno da média da população. Dessa forma, se torna mais intuitivo a escolha da amostra adequada.

No que concerne aos dados geoespaciais, devemos observar também se a amostra representa as variações existentes quanto ao tempo e ao espaço. Quanto às variações temporais, o conjunto de dados pode cobrir um intervalo de tempo que compreenda feriados, períodos de férias e até diferentes estações do ano. Tendo em vista as mudanças na rotina dos usuários, é preciso considerar na amostragem a separação de intervalos de tempo que possam necessitar de análises específicas. Por sua vez, a variação espacial deve ser considerada devido a possíveis diferenças no comportamento dos indivíduos em cada região. Por exemplo, se o conjunto possui dados em 10 estados brasileiros, devemos nos certificar que a amostra coletada também possua dados em todos os 10 estados e não apenas em um. Se aplicável, pode-se também produzir amostras de cada região analisada separadamente, apresentando os resultados para cada uma.

Existem algumas técnicas relacionadas à amostragem orientada a dados geoespaciais, como a amostragem simples orientada a feições ou a área. Tais abordagens têm como propósito garantir a qualidade das amostras geradas [Tong et al. 2011] e, portanto, não são focadas na análise de dados propriamente dita.

3.5.1.5. Agregação

Com os dados já prontos para utilização, devemos analisar se as informações que obtivemos já são suficientes para o propósito esperado, pois pode ser que uma última etapa

para enriquecimento dos dados seja necessária. Esta etapa visa agregar dados provenientes de outras fontes aos dados originais. Informações acerca de setores censitários, clima, criminalidade, dentre outros, podem ser incorporadas à sua aplicação, sendo que o seu uso dependerá unicamente dos objetivos do projeto. Entretanto, o uso exacerbado de tais fontes pode acarretar em problemas relacionados à memória e ao tempo de resposta. Vale lembrar ainda, que a alta dimensionalidade dos dados, em casos relacionados ao aprendizado de máquina, pode levar o modelo a não generalizar bem, tornando seu uso pouco atrativo em casos reais. Portanto é sempre bom verificar quais variáveis agregadas realmente fazem a diferença e quais podem ser descartadas.

Para realizar a agregação dos próprios dados com os de outras fontes, podemos utilizar a mesma estratégia adotada durante a filtragem, identificando primeiramente a representação geométrica de ambas as representações. Se nossos dados estão no formato de pontos e queremos adicionar informações da região demográfica onde cada um está localizado, devemos obter os polígonos de cada região, com as devidas características de cada uma e, para cada ponto, verificar em qual polígono ele está contido. Apesar desta estratégia poder ser utilizada para a maior parte das agregações baseadas em informações geográficas, devemos nos atentar ao desempenho, lançando mão de abordagens como a indexação, discutida anteriormente. Se usarmos como exemplo os setores censitários mapeados pelo IBGE, o custo de se mapear milhares de pontos em um dos mais de 300 mil setores é uma operação muito custosa computacionamente.

Outro caso possível é a agregação de dados com representação de pontos ou linhas. Neste cenário, podemos utilizar a operação de *BUFFER* para obtermos um polígono que representa a área daquela região com uma margem adequada ao propósito. Por exemplo, para uma aplicação relacionada ao tráfego com o intuito de indicar ao usuário qual região tem trânsito pesado. Neste caso, ao localizarmos pontos referentes ao usuários, dificilmente estes estarão exatamente sobre a linha que representa a via, sendo necessário considerar uma pequena margem.

Vale lembrar também que várias agregações são possíveis ao se analisar o próprio conjunto de dados como, por exemplo, inferências de laços sociais, localização de locais como casa e trabalho do usuário, identificação de pontos de interesse, dentre outros. Essas agregações envolvem algoritmos e implementações que podem ser integradas ao projeto ou desenvolvidas com o propósito específico.

3.5.2. Extração de Conhecimento

A seguir, introduzimos algumas das aplicações possíveis para a extração de conhecimento de dados geoespaciais. Para cada uma, apresentamos exemplos de seu funcionamento e destacamos a sua importância na análise de dados.

Para alguns dos tópicos abordados, o cálculo da distância entre duas coordenadas requer um cuidado especial. Isso porque o formato elipsoidal da Terra faz com que métodos como o da distância Euclidiana contenham erro que pode ser inconcebível. Nesta situação, outros dois métodos se destacam: a fórmula de Haversine e a fórmula de Vincentys. Apesar dos erros, os três métodos podem ser utilizados, dependendo da aplicação.

O fato de o método Euclidiano inferir que a distância entre dois pontos é uma

reta, faz com que este tenha o maior erro dentre os três. Isso porque, para calcular a distância entre coordenadas distantes, este erro se agrava, impactando negativamente a aplicação. Por outro lado, se os dados utilizados foram projetados em um plano, ou se a distância esperada é muito pequena, a distância Euclidiana pode ser utilizada. Já a fórmula de Haversine considera a distância entre dois pontos como uma curva, fazendo com que seja mais adequado ao formato da Terra. Entretanto, por considerar a Terra como uma esfera e não uma elipse, este método também apresenta erros, apesar de mais brandos, fazendo com que o método seja um dos mais utilizados em aplicações para dados georreferenciados. Por último, o método de Vincentys calcula a distância entre dois pontos com base em uma elipse, fazendo com que este seja o método mais acurado dentre os três. Por outro lado, este é um método computacionalmente custoso.

3.5.2.1. Raio de Giro

Quando trabalhamos com dados geoespaciais, é comum assumirmos que os usuários analisados possuem uma localização de referência, que pode ser sua casa, seu local de trabalho ou algum outro ponto de interesse. Com base nesse princípio, podemos calcular o raio de giro de um usuário, que pode ser definido como a distância máxima entre a localização de referência da entidade e as outras localidades visitadas por ela. O raio de giro fornece informações quanto à mobilidade dos usuários, permitindo, por exemplo, a classificação das entidades quanto ao valor de seus raios. A Figura 3.11 apresenta os passos para o cálculo do raio de giro de um usuário a partir de suas localizações reportadas, destacando a identificação do ponto de referência e o cálculo das distâncias.

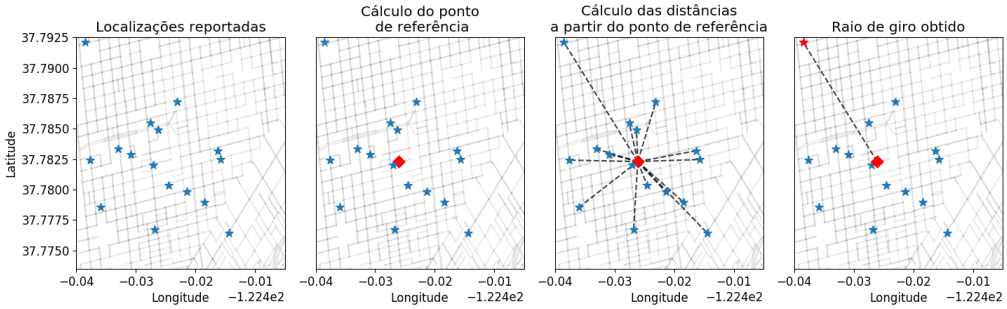


Figura 3.11. Cálculo do raio de giro de um usuário com base nas suas localizações reportadas.

Para calcular o raio de giro de um usuário, é preciso definir primeiramente o critério para a identificação da sua localização de referência através dos dados geoespaciais. Algumas abordagens encontradas na literatura são a escolha de um ponto aleatório [Kosta et al. 2012], a média de todos os pontos reportados, o local mais visitado e o primeiro local reportado no dia [Ekman et al. 2008]. É preciso levar em conta as características do conjunto de dados, como a granularidade das amostras e os sensores usados, para escolher um critério que identifique localizações de referência que façam sentido. Adicionalmente, deve-se escolher uma função de distância (como haversine ou euclidiana) para realizar o cálculo entre os pontos.

Conhecer o raio de giro de usuários tem ajudado em pesquisas sobre mobilidade

urbana já que, uma vez que espera-se que com o passar do tempo uma pessoa explore áreas cada vez maiores [González et al. 2008], permitindo assim antecipar ações necessárias. Análises de mobilidade em contextos mais específicos, como locais visitados por usuários de redes sociais, como o Twitter por exemplo, também são possíveis [Jurdak et al. 2015]. Além disso, o uso do raio de giro auxilia em diversas outras áreas como, por exemplo, o entendimento de rotas de fuga durante desastres naturais [Wang e Taylor 2014] e estudos de alcance de ações aplicadas mundialmente [Morales et al. 2017].

3.5.2.2. Agrupamentos Espaciais

Outra atividade comum ao lidar com dados geoespaciais é a necessidade de se identificar grupos que possuam padrões de comportamento semelhantes. Tais padrões podem ser representados por usuários que frequentam a mesma região [Sakai et al. 2014], identificação de regiões de maior demanda de serviços, locais com focos de disseminação de uma doença, dentre outros.

Existem diversos algoritmos de agrupamento na literatura, utilizando-se de diversas abordagens para obter os grupos, como algoritmos com base em distância (*K-means*), em densidade (*DBScan*) e em distribuição (*GMM*). Apesar de serem algoritmos amplamente utilizados para conjuntos de dados comuns, o seu uso indiscriminado em dados referentes à localização pode gerar erros. Algoritmos baseados em distância muitas vezes trabalham com distância euclidiana, que não representa bem a distância entre pontos no globo terrestre. Isso porque a distância euclidiana comporta-se melhor quando aplicada sobre planos e não sobre uma forma elipsoide como a da Terra [Ingole e Nichat 2013]. Uma medida de distância adequada, apesar de não perfeita, para este tipo de agrupamento seria a distância de *Haversine*, pois considera a distância entre dois pontos como um arco. Já os algoritmos baseados em distribuição dependem da inferência de que os dados seguem algum tipo de distribuição, podendo não ser verdade para o conjunto de dados.

Assim, os algoritmos mais populares para agrupamento espacial seriam os baseados em densidade, já que esta pode ser aplicada para os mais diversos dados. Entretanto, estes algoritmos têm de ser bem ajustados para que os grupos sejam representativos. Como vemos na Figura 3.12, os grupos obtidos com a utilização dos algoritmos *DBScan* e *K-Means* diferem entre si. Isso mostra que a utilização de determinado algoritmo dependerá dos dados e sua projeção, além da necessidade da aplicação, já que nem sempre teremos as condições ideais para uma escolha clara.

3.5.2.3. Pontos de Interesse

Os pontos de interesse (PoIs) representam localizações visitadas pelas entidades com certa frequência. Esses locais podem ser restaurantes, estacionamentos, escritórios de trabalho, dentre outros. Detectar os pontos de interesse existentes em uma região é tarefa fundamental para melhor compreender a mobilidade dos indivíduos, pois permite identificar a motivação, a frequência e a duração das visitas. Com essa informação em mãos, redes podem ser ajustadas para atender a demanda necessária, sistemas de recomendação podem sugerir locais semelhantes àqueles visitados, e campanhas de marketing e publicidade

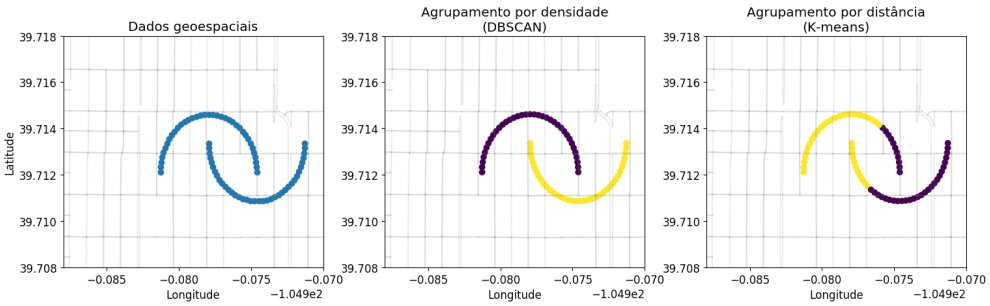


Figura 3.12. Dados geoespaciais podem ser agrupados de acordo com as relações existentes entre os pontos, como densidade e distância.

podem ser personalizadas de acordo com o local visitado [Tran et al. 2013].

Existem vários algoritmos para a detecção de PoIs, os quais utilizam diferentes estratégias de detecção dependendo do tipo de dado geoespacial e características da coleta, como granularidade. Os dados de *check-in* e LBSN, por exemplo, são comumente reportados com dados de PoIs, necessitando pouco ou nenhum processamento. No caso de dados de sensores GPS, faz-se necessário o processamento dos pontos, considerando a formação de grupos em torno de áreas de raio reduzido (Figura 3.13). Nesse cenário, algoritmos de agrupamento espacial – como o DBSCAN e o GMM – são capazes de produzir resultados satisfatórios. Por fim, pode-se também realizar fusões de dados para adicionar ao conjunto de dados analisado a localização dos PoIs em uma região.

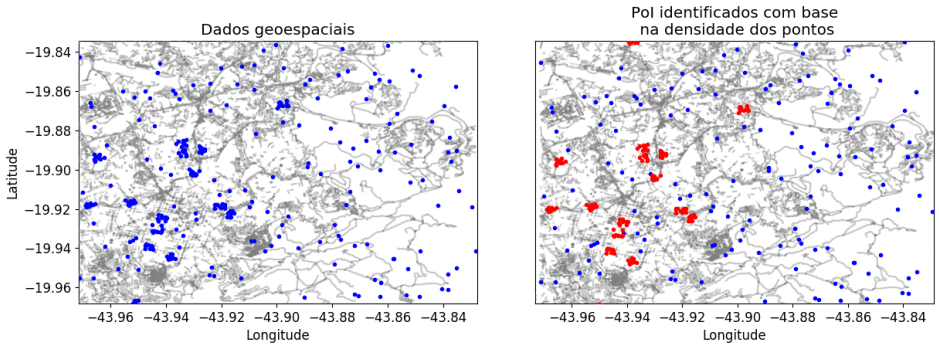


Figura 3.13. Detecção de pontos de interesse (PoIs) a partir de dados geoespaciais

3.5.2.4. Características sociais

A inferência de relações sociais entre indivíduos de uma base de dados pode nos auxiliar a entender melhor diversas atividades do nosso cotidiano, como padrões de mobilidade que se alteram devido a outras pessoas [Cho et al. 2011], análise da adesão a programas de prevenção [Choi et al. 2017] ou ainda a possibilidade de se projetar a propagação de doenças em uma epidemia [Firestone et al. 2011]. Para que isso seja possível, a identificação do laço social correto entre dois usuários é de vital importância, fazendo com que tenhamos cuidado ao decidir quais tipos de relação iremos mapear.

Os laços sociais derivados de uma análise podem dizer se duas pessoas são amigas, conhecidas, vizinhas, colegas de trabalho ou se moram juntas, por exemplo. Porém, a inferência de tais laços não é uma atividade simples, pois não conseguimos medir com clareza quantos encontros entre usuários são necessários para que estes evoluam de desconhecidos para amigos, por exemplo. Para este fim, uma abordagem bastante utilizada é a construção de grafos de contatos, onde cada vértice representa um usuário e uma aresta indica o encontro entre eles (as arestas podem conter pesos para identificar quantas vezes os usuários se encontraram ou a duração dos encontros, por exemplo).

No que compete aos dados geoespaciais em si, existem trabalhos na literatura que utilizam as informações de grafos de contatos para identificar probabilidades de usuários possuírem gostos em comum através da ocorrência de rotas similares [Hung et al. 2009]. Já outros trabalhos também analisam a ocorrência de comunidades, porém a partir de informações espaço-temporais [de Melo et al. 2015, Li et al. 2008].

3.6. Visualização

A visualização de dados geoespaciais é parte essencial em todas as etapas de sua utilização em uma análise. Ela começa a ser aplicada logo após a coleta, para fins de validação e verificação dos dados coletados, onde falhas podem ser detectadas e corrigidas através de novas coletas ou etapas de pré-processamento. Em seguida, diversas visualizações são criadas para retratar a distribuição das informações, permitindo a detecção de *outliers* e o mapeamento das características da população analisada. Durante o desenvolvimento da análise, as visualizações auxiliam nas tomadas de decisão e na apresentação de resultados parciais. Por fim, os resultados finais da análise também são apresentados através de visualizações, que facilitam o compreendimento das ideias propostas.

Sendo assim, é fundamental que os gráficos criados sejam de fácil leitura e que sua construção não demande tempo considerável. Esses dois fatores garantem que a utilização de gráficos seja uma solução para o processo, e não mais um problema a ser resolvido. Para que isso aconteça, é preciso conhecer as diferentes formas de visualização dos dados geoespaciais e os resultados que elas oferecem. Diferente de dados numéricos e categóricos, em que gráficos tradicionais como de Barra, Pizza, Linhas, dentre outros, são suficientes para transmitir o conteúdo, a visualização de dados geoespaciais geralmente envolve a necessidade de um mapa sobre o qual as localizações serão desenhadas. Adicionalmente, precisamos considerar estratégias de agregação devido ao grande volume de dados. Desenhar quantidades massivas de dados pode não ser eficiente computacionalmente, além de produzir resultados poluídos. Além disso, por vezes estamos mais interessados na visualização dos padrões existentes nos dados do que no comportamento de indivíduos específicos.

Nesta seção, discutimos sobre as técnicas e ferramentas necessárias para gerar visualizações eficazes. Primeiramente, a Seção 3.6.1 apresenta as formas possíveis de visualização de dados geoespaciais, suas vantagens e desvantagens, e quais critérios levar em consideração para escolher uma forma. Em seguida, a Seção 3.6.2 discute algumas ferramentas que facilitam a construção destas visualizações.

3.6.1. Tipos de visualização e suas propriedades

Os tipos de visualização apresentados, a seguir, são as estruturas básicas para a construção de gráficos para a análise de dados geoespaciais. A partir dessas abordagens, é possível desenvolver novas visualizações, adaptando e adicionando características para atender aos requisitos da visualização desejada. Ao realizar os primeiros esboços, o leitor notará que modificações serão necessárias para tornar o gráfico proposto o mais claro possível, o que é essencial para o seu entendimento. Assim, podemos citar o formato do mapa, o nível de *zoom*, a escala de cores usadas, legendas visuais e textuais, e a localização central.

Para auxiliar no entendimento dos gráficos de dados geoespaciais, utilizamos os mapas das regiões representadas para aproximar a figura do ambiente real. Para desenhar um mapa, devemos primeiro definir o seu tipo e a sua projeção. O tipo do mapa define as características que ele apresenta, como relevo, divisas territoriais e ruas e rodovias. Já a projeção compreende à representação dos dados em duas ou três dimensões. Ambos os fatores devem ser escolhidos de forma a fornecer um melhor entendimento do gráfico gerado, ao mesmo tempo em que características não essenciais devem ser desconsideradas para evitar a sua poluição. Por exemplo, desenhar um conjunto de dados geoespaciais de duas dimensões em uma projeção de três dimensões não adicionará informação à figura, podendo inclusive gerar incertezas quanto ao seu significado.

O nível de *zoom* se refere à ampliação da região coberta pela figura, podendo se aproximar (*zoom in*) ou se distanciar (*zoom out*). Um gráfico deve cobrir toda a área em que os dados geoespaciais estão localizados, a menos que o objetivo seja enfatizar alguma região específica. Porém, na presença de *outliers* podemos gerar visualizações afastadas da grande região de interesse e, assim, nesses casos a aproximação pode ser considerada.

A escala de cores de um gráfico não possui função apenas estética, podendo ser usada para auxiliar na indicação de níveis de intensidade (como altitude) e na separação de classes existentes nos dados (como tipo de rede ou dispositivo). Quando for o caso, deve-se optar pela escolha de escalas de cores com alto contraste (de fácil identificação também ao visualizar a figura em escala de cinza) e que remetam aos valores das classes representadas (por exemplo, usar tons de azul e vermelho para indicar a temperatura). Na presença de um número elevado de classes ou de escalas de variação complexas, as escalas de cores podem produzir visualizações confusas e, portanto, podem ser substituídas por legendas textuais (valores escritos ao invés de indicados por cores) e visuais (uso de símbolos para separar as categorias de dados).

3.6.1.1. Ponto-a-ponto

O gráfico ponto-a-ponto constitui a forma mais simples de visualização de dados geoespaciais. Nele, os dados são desenhados como pontos sobre o mapa de acordo com as suas coordenadas. Os pontos podem ter diferentes tamanhos, cores e formatos para retratar diferentes características dos dados. Por um lado, é um tipo de visualização facilmente compreensível, de implementação e modificação simples (Figura 3.14(a)). Por outro lado, a visualização de grandes quantidades de dados em gráficos ponto-a-ponto causa a sobreposição dos pontos (Figura 3.14(b)) o que leva a perda de detalhes, sem mencionar o consumo de memória para que seja construído. Sendo assim, os gráficos ponto-a-ponto

devem ser utilizados para visualizar a dispersão dos dados coletados, e evitados quando se deseja explorar os detalhes presentes no conjunto.

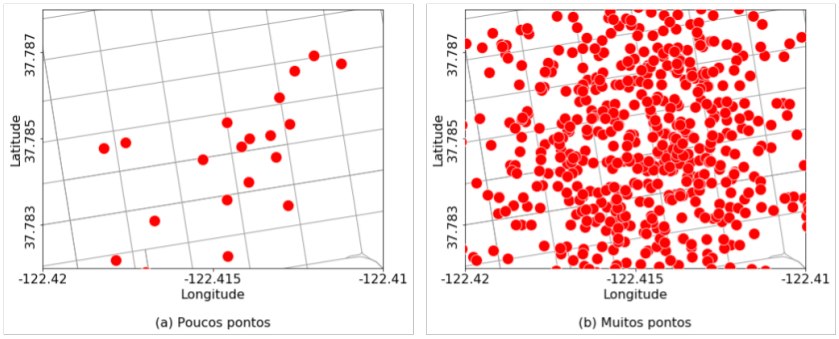


Figura 3.14. O gráfico ponto-a-ponto para poucos dados permite uma distinção clara entre os pontos (a); o mesmo não ocorre para um conjunto de muitos dados (b), onde é infactível obter uma separação entre os pontos desenhados.

3.6.1.2. Mapa de calor

Os gráficos de mapa de calor representam a densidade de uma variável dos dados geo-espaciais através de curvas de intensidade e escalas de cores. Isto leva à formação de regiões de alta densidade, que podem representar uma topologia ou pontos de interesse, por exemplo (Figura 3.15(a)). Os gráficos de mapa de calor são recomendados quando os dados não possuem uma distribuição de localizações uniforme, o que leva a existência de regiões de maior intensidade. Por serem baseados na densidade dos pontos, os gráficos de mapa de calor tendem a ser pouco informativos para conjuntos de dados de localização esparsa (Figura 3.15(b)).

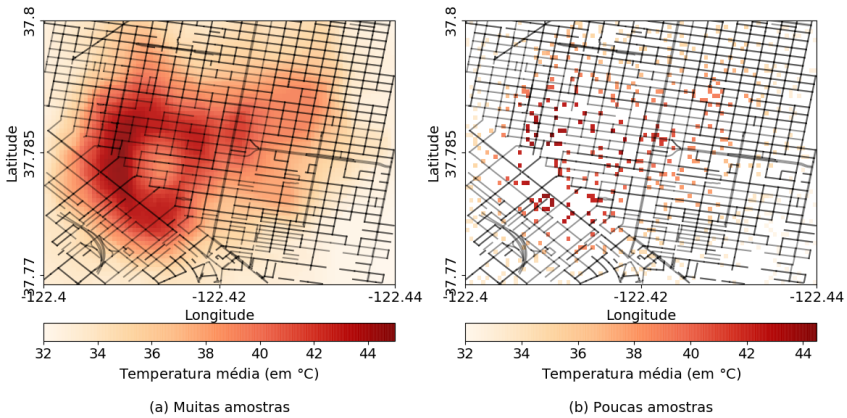
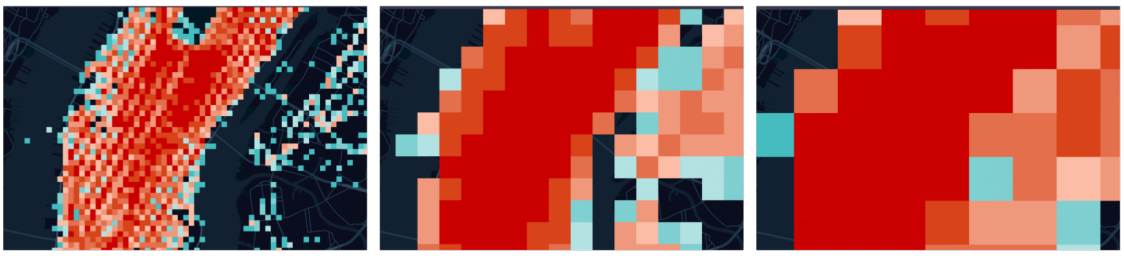


Figura 3.15. Os mapas de calor permitem detectar variações na intensidade dos pontos (a), porém perdem sua utilidade se o volume de dados coletados é esparso (b).

3.6.1.3. Grade

A visualização em grade é feita através da divisão da região analisada em uma grade com dimensões definidas, onde a unidade mínima de posição são as células que compõem a grade. Para cada célula, os dados localizados nas coordenadas contidas dentro dela são agregados através de uma função de agregação (por exemplo, a soma, a média, o valor mínimo ou o valor máximo) e o resultado da função representa a célula. Além da função de agregação, outro parâmetro a ser definido é o tamanho da célula, isto é, a área coberta por ela. Células menores são capazes de exibir mais detalhes (Figura 3.16(a)) enquanto células maiores perdem informações que podem ser relevantes (Figura 3.16(c)). Eventualmente, uma solução de compromisso pode ser necessária (Figura 3.16(b)).



(a) Escala pequena (a) Escala média (a) Escala grande

Figura 3.16. A visualização através de grade permite observar as variações na densidade dos pontos (e também de outras informações presentes nos dados), porém mudanças no tamanho da grade ocasionam na perda de detalhes devido ao agregamento.

3.6.1.4. Fluxo

O gráfico de fluxo tem como papel representar o fluxo de entidades (que podem ser pessoas, veículos ou outros) entre duas ou mais regiões. Esse deslocamento é representado por arcos que conectam as regiões de origem e de destino, juntamente com a indicação de sua intensidade, que pode ocorrer através das dimensões dos arcos (e.g., um arco mais largo indica um fluxo maior) ou através de escalas de cores. É válido ressaltar que os arcos podem não ser simétricos, isto é, a intensidade do fluxo a partir da origem até o destino pode ser diferente daquela obtida partindo do destino até a origem. Sendo assim, múltiplos gráficos podem ser desenhados para que todos os casos sejam cobertos.

Existem outras formas encontradas na literatura para representar fluxos entre regiões, como grafos de transição e matrizes de transição. Em relação ao primeiro, os nós representam as regiões, as arestas representam uma conexão entre regiões e o peso da aresta representa a intensidade desse fluxo (Figura 3.17). De maneira similar, as matrizes de transição contêm em seus eixos as regiões analisadas, e os termos individuais representam o fluxo correspondente à linha e à coluna aos quais o termo está contido.

Figura 3.17. Os gráficos de fluxo representam a intensidade das interações entre regiões diferentes. Para isso, conectam arcos entre as regiões de origem e de destino e utilizam de elementos visuais (como as dimensões da curva, etiquetas textuais ou escalas de cores) para indicar sua intensidade.

3.6.1.5. Em barras

Estes gráficos utilizam uma terceira dimensão para representar barras que indicam o impacto de uma variável sobre regiões de tamanho definido que compõem a área analisada. Assim como os gráficos em grade, são úteis quando os dados geoespaciais possuem informações sensíveis à localização. Por sua vez, estes gráficos permitem observar regiões de destaque com maior facilidade, devido à projeção das barras em uma terceira dimensão. Porém, como pode ser notado no exemplo de gráfico da Figura 3.18, é preciso considerar os limites de se representar dados em três dimensões através de imagens não interativas. Devido à projeção das barras, algumas regiões podem deixar de ser visíveis, causando perda de informação.

3.6.2. Ferramentas

Nesta seção, serão discutidas as principais bibliotecas e ferramentas existentes para visualização de dados geoespaciais. Esses recursos facilitam a aplicação das visualizações apresentadas na Seção 3.6.1 pois proveem interfaces para leitura dos dados, criação e personalização dos gráficos, e sua exportação para diversos formatos. Por fim, ressaltamos que as bibliotecas e ferramentas apresentadas são de uso livre e possuem código aberto e, portanto, podem ser aplicadas em análises sem custo e alteradas de acordo com as necessidades de uso.

3.6.2.1. Bokeh

Bokeh [Bokeh Development Team 2019] é uma biblioteca para construção e visualização de gráficos e figuras. Ela tem uma interface Web e sua implementação é feita na lingua-

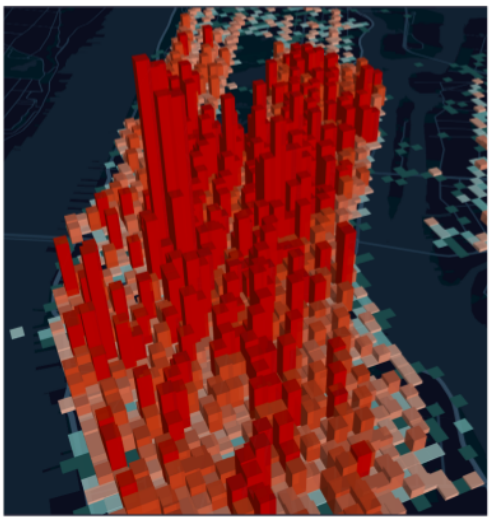


Figura 3.18. Os gráficos em barras são uma versão em três dimensões dos gráficos em grade, permitindo visualizar as variações existentes entre regiões demarcadas. Entretanto, é preciso considerar que a sua visualização em duas dimensões pode ser prejudicada pelas estruturas presentes.

gem Python. A sua principal vantagem é a interatividade de suas visualizações, permitindo ao usuário alterar valores e escalas, e inserir novos dados em um gráfico existente, tudo feito em tempo real. Isso faz com que a biblioteca seja uma alternativa interessante para a publicação de resultados de análises em websites. Outro ponto é o suporte para a carga de grandes volumes de dados, que podem ser importados através da biblioteca *Pandas*. Por fim, destaca-se também a extensa coleção de visualizações disponíveis para uso, que contemplam diversas áreas de estudo de dados comuns e geoespaciais.

3.6.2.2. Kepler

Kepler [Vis.gl 2018] é uma ferramenta de análise de dados geoespaciais desenvolvida pela *Uber*. Ela possui interface Web de uso simples, que permite carregar os dados geoespaciais desejados, realizar agregações e filtragens de dados, e projetá-los sobre um mapa detalhado do globo terrestre usando diversas visualizações diferentes, como ponto-a-ponto, mapas de calor, gráficos de fluxo e de barra. Além disso, o usuário pode escolher a projeção do mapa (duas ou três dimensões), desenhar figuras geométricas para auxiliar na construção das visualizações e exportar as figuras para diversos formatos. Porém, por utilizar um mapa de coordenadas latitude e longitude para prover suas visualizações, a ferramenta não possibilita que outros tipos de coordenadas sejam usadas como entrada.

3.6.2.3. Leaflet

Leaflet [Agafonkin 2010] é uma biblioteca para a construção de mapas interativos e *mobile-friendly* desenvolvida na linguagem Javascript. Ela usa os dados do *OpenStreetMaps* para construir a projeção de mapas detalhados contendo informações de vias e

demarcações de locais e transportes públicos. Os mapas criados possuem funções interativas de visualização e edição, podendo também enviar e receber informações em tempo real. Assim como Kepler, Leaflet só é capaz de criar visualizações de dados que utilizam coordenadas de latitude e longitude.

3.6.2.4. OSMnx

OSMnx [Boeing 2017] é uma biblioteca implementada em Python para visualização de dados geoespaciais focada em mapas rodoviários. Usando dados vindos do *OpenStreetMaps*, ela é capaz de gerar visualizações personalizadas da malha viária de uma determinada região desejada, sobre a qual o usuário pode desenhar seus dados geoespaciais. Além disso, ela constrói a rede das vias através de grafos, permitindo o mapeamento de pontos às vias mais próximas, o cálculo de distâncias e caminhos mínimos considerando as vias, e a aplicação de métricas e métodos de redes complexas.

3.6.2.5. QGis

QGis [QGIS Development Team 2009] é um software multiplataforma que permite a visualização e a edição de dados georreferenciados para análise. É um sistema robusto, capaz de carregar grandes volumes de dados e que suporta diversos tipos de dados de entrada. Devido às suas capacidades avançadas, consegue produzir visualizações com alta qualidade. Por outro lado, isso implica em uma curva de aprendizado maior. Finalmente, a sua disponibilidade como uma ferramenta "isolada" torna complexa a sua interação com os demais processos da análise de dados geoespaciais, o que pode desfavorecer o seu uso para a construção de visualizações rápidas.

3.7. Tópicos em Aberto

Conforme apresentado anteriormente na Figura 3.1, o campo de pesquisa em dados geoespaciais ainda é recente e muito mais precisa ser feito. Nesse cenário, considerando as etapas que compõem o processo de análise de dados geoespaciais, existem linhas de pesquisa pouco exploradas. Dessa maneira, com o objetivo de investigar as questões em aberto, apresentamos nesta seção alguns dos desafios existentes referentes aos mecanismos de privacidade para anonimização de dados, detecção e classificação de pontos de interesse, preenchimento de lacunas espaciais e temporais nos dados e a fusão de dados heterogêneos. Esperamos que, com as informações e as referências introduzidas, o leitor possa vislumbrar potenciais projetos de pesquisa e desenvolvimento.

3.7.1. Mecanismos de privacidade

A análise de dados geoespaciais envolve a extração de conhecimento das localizações reportadas pelos usuários, as quais podem ser aplicadas na detecção de pontos de interesse, no mapeamento de fluxos de mobilidade e na predição da demanda de recursos de rede, por exemplo. Essas aplicações fazem uso das informações obtidas dos dados agregados e, portanto, não necessitam do conhecimento do comportamento individual de cada usuário. Porém, ao compartilhar dados geoespaciais é necessário considerar a presença de agen-

tes mal-intencionados – denominados atacantes – que desejam utilizar os dados para fins próprios que violam a privacidade dos usuários – denominados vítimas – colocando os mesmos em risco [Krumm 2009]. Como exemplo da atuação de atacantes, podemos citar os ataques de re-identificação pessoal, de identificação de pessoas próximas (através da análise de localizações e rotinas), de identificação de pontos de interesse, de rastreamento da localização em tempo real e de confirmação da presença (ou ausência) do indivíduo em um determinado local e momento. Com essas informações em mãos, o atacante pode planejar ações na vida real, como roubos e sequestros.

Portanto, é fundamental garantir a segurança dos usuários sensoreados sem que haja a interrupção do compartilhamento dos dados geoespaciais, os quais são essenciais para o funcionamento de diversos sistemas baseados em localização. Para isso, existem os mecanismos de proteção à privacidade de localização (*Location Privacy Protection Mechanisms* ou LPPM) que atuam na anonimização dos dados geoespaciais sem que suas características sejam perdidas. É válido ressaltar que existem LPPMs específicos para prevenir diferentes tipos de ataque, sendo necessário, assim, a composição de LPPMs quando há a necessidade de prevenir múltiplos ataques. Por outro lado, o uso de múltiplos LPPMs pode acarretar na perda da qualidade dos dados, tornando ineficaz o conhecimento gerado.

A utilização de LPPM para a anonimização de dados geoespaciais é um conceito recente na literatura, o que implica na dificuldade de reprodução dos trabalhos apresentados. Um problema recorrente é a falta de padronização no que diz respeito às definições formais de LPPMs e seus possíveis ataques. Isso dificulta a comparação entre os mecanismos, devido a diferenças quanto aos dados usados como entrada e produzidos como saída e quanto as métricas e métodos utilizados para avaliação da qualidade dos resultados produzidos. Outro problema recorrente compete na difícil utilização de LPPMs encontrados na literatura, visto que possuem alta complexidade de implementação e aplicam muitas vezes técnicas de aprendizado profundo, as quais demandam grande volume de dados. Por fim, destacamos também a necessidade de explorar as variações do compromisso entre a qualidade dos resultados obtidos pelos dados anonimizados e o nível de privacidade preservado.

3.7.2. Detecção de Pontos de Interesse

A detecção de pontos de interesse é questão fundamental no estudo da mobilidade humana e sua evolução com o tempo. Pontos de interesse em dados geoespaciais podem representar residências, lojas, escritórios de trabalho e até semáforos e regiões de congestionamento frequente. Para isso, é necessário desenvolver algoritmos que entendam os padrões de movimentação dos usuários para extrair os pontos de interesse desejados. Tais padrões podem ser modelados de acordo com os tempos mínimos e máximos de permanência em um local, o seu raio de cobertura e a frequência de visitação. Esses parâmetros variam de acordo com as entidades analisadas (humanos ou veículos) e com os costumes culturais – padrões de rotina e meios de transporte – da região analisada.

Podemos destacar dois desafios quanto à extração de pontos de interesse. O primeiro é a anonimização dos dados geoespaciais, que ocorre, como visto na Seção 3.7.1, devido à necessidade de proteger as informações latentes dos usuários sensoreados. Ao

utilizar dados anonimizados, o conjunto de localizações de um mesmo usuário (identificável antes da etapa de anonimização) é dividido em identificadores diferentes e não relacionados. Devido a isso, a contagem da frequência de visitas de um usuário para a detecção de seus pontos de interesse se torna infactível, levando à necessidade de desenvolver novos métodos para tal. Nesse caso, é possível identificar pontos de interesse coletivos, como pontos turísticos, que são comuns a vários usuários.

O segundo desafio quanto à detecção de pontos de interesse é a escassez de dados, seja devido a uma quantidade reduzida de usuários ou à existência de lacunas espaciais e temporais na coleta. No primeiro caso, é possível agregar mais usuários à região analisada através da fusão de dados. Quanto ao segundo, é preciso a aplicação de algoritmos de detecção que considerem a existência de lacunas, inclusive considerando-as como possíveis pontos de interesse. Adicionalmente, é possível realizar o preenchimento das lacunas existentes, resultando em trajetórias mais detalhadas e completas no tempo e espaço, tornando os resultados da detecção mais completos e acurados.

3.7.3. Preenchimento de lacunas

Como visto anteriormente, a capacidade dos sensores e as estratégias de coleta de dados adotadas podem levar à existência de lacunas espaciais e temporais nos dados. Essas lacunas causam uma caracterização errônea da mobilidade dos usuários e suas interações, afetando os resultados das análises baseadas nessas informações [Silva et al. 2015, Cunha et al. 2016].

As abordagens atuais para preenchimento de lacunas espaciais e temporais podem ser divididas em dois grupos: aquelas que usam métodos puramente matemáticos, como interpolação e extrapolação [Hoteit et al. 2016], e aquelas que usam algoritmos baseados em dados históricos [Chen et al. 2019, Celes et al. 2017]. Enquanto as soluções do primeiro grupo são capazes de produzir resultados satisfatórios para lacunas pequenas, suas estimativas são pouco precisas para grandes lacunas. Por sua vez, as estratégias do segundo grupo tendem a obter melhores resultados em lacunas maiores, porém ao custo da necessidade de dados históricos nem sempre disponíveis.

Quando consideramos dados de extrema esparsidade, como trajetórias em que somente os pontos de origem e destino estão disponíveis, ainda são poucas as abordagens capazes de gerar resultados [Domingues et al. 2018]. Nesse caso, é preciso definir as possíveis rotas a serem tomadas, a velocidade na qual elas serão percorridas, entre outros fatores.

3.7.4. Fusão de dados heterogêneos

Com os crescentes avanços no desenvolvimento de aplicações para a computação urbana, torna-se cada vez mais necessário o sensoriamento e a coleta de dados de diversas fontes diferentes, como mídias sociais, tráfego, mobilidade humana, meteorologia, pontos de interesse e outros [Zheng et al. 2014]. Mais do que servirem a propósitos isolados, os diferentes tipos de dados serão combinados e aplicados na construção de sistemas ubíquos e cientes de contexto, capazes de fornecer um vasto conjunto de serviços aos usuários, como rotas veiculares otimizadas que consideram o tráfego e o clima e recomendações personalizadas de pontos de interesse.

Para que isso ocorra, métodos de fusão de dados precisam ser utilizados. A fusão de dados pode ser homogênea, isto é, quando os dados se assemelham em suas características essenciais (e.g., dois conjuntos de dados de trajetórias de táxis), ou heterogênea, onde os conjuntos de dados descrevem eventos diferentes (e.g., um conjunto de trajetórias de táxis e um relatório de índices de tráfego). Enquanto que a fusão de dados homogênea possui implementação simples, a fusão de dados heterogênea é de complexidade elevada devido às diferentes dimensões dos dados e a assincronia dos eventos [Rettore et al. 2020]. Até o presente momento, não existem abordagens consolidadas para a fusão heterogênea de dados geoespaciais que apresentem os conceitos, métodos e ferramentas necessários.

3.8. Conclusão

Este capítulo apresentou um estudo aprofundado sobre dados geoespaciais e sua aplicação na extração de conhecimento e geração de valor por meio de novos produtos e serviços, permitindo a captação de novas receitas, além de avançar o estado da arte em áreas relacionadas à mobilidade, internet das coisas, computação urbana, dentre outras. Para isso, foram apresentados os conceitos teóricos e as principais técnicas e ferramentas aplicadas às etapas de coleta, armazenamento, tratamento, extração de conhecimento e visualização de dados geoespaciais.

Inicialmente, foi destacada a importância da mobilidade para o desenvolvimento de novas técnicas e tecnologias aplicadas a questões como controle e predição de fluxo de trânsito, modelos de contágio, otimização de recursos de rede, dentre outras. Nesse cenário, o interesse pelo estudo e a aplicação de dados geoespaciais têm sido cada vez maior, devido à sua capacidade de representar o comportamento de mobilidade de entidades, tais como seres humanos e veículos. Porém, apesar da crescente demanda por pesquisas envolvendo esse tipo de dado, ainda não há um consenso em relação às metodologias adequadas para sua análise, devido à escassez de referências que apresentem, de forma clara, os conceitos e técnicas a serem aplicados.

Dessa maneira, visando preencher essa lacuna, a Seção 3.2 introduziu os principais conceitos relacionados aos dados geoespaciais, como as características geográficas da Terra, os sistemas de referência, as projeções espaciais e a manipulação numérica de coordenadas. A Seção 3.3 discutiu o processo de coleta de dados, com destaque para as principais fontes existentes (dispositivos GPS, *smartphones* e redes sociais baseadas em localização). Além disso, foram destacadas as características inerentes aos dados coletados, como privacidade das entidades sensoreadas, granularidade, acurácia e precisão. Em seguida, a Seção 3.4 apresentou as particularidades do armazenamento de dados geoespaciais, os quais necessitam de sistemas de gerenciamento de bancos de dados, formas de representação, métodos de compactação e estruturas de indexação específicos para serem armazenados de forma eficiente. A Seção 3.5 apresentou as etapas de preparação dos dados e a extração de conhecimento. Na preparação, destacam-se as sub-tarefas de formatação, amostragem, limpeza, filtragem e agregação dos dados. Na etapa de extração de conhecimento foram detalhadas algumas aplicações, como o cálculo do raio de giro e a detecção de agrupamentos espaciais, pontos de interesse e características sociais. As técnicas de visualização de dados, juntamente com as bibliotecas e ferramentas utilizadas para a sua criação foram discutidas na Seção 3.6. Por fim, a Seção 3.7 discutiu algumas

questões em aberto no estudo e na aplicação de dados geoespaciais, com o intuito de elevar o interesse do leitor quanto a potenciais projetos de pesquisa e desenvolvimento.

Agradecimentos

Este trabalho foi realizado com apoio da CAPES, CNPq, Cinnecta e Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), processos 15/24494-8 & 18/23064-8.

Referências

[mys 2020] (2020). MySQL :: MySQL 8.0 Reference Manual :: 11.4 Spatial Data Types. [Online; accessed 15. Apr. 2020].

[ogc 2020] (2020). The Home of Location Technology Innovation and Collaboration | OGC. [Online; accessed 19. Apr. 2020].

[Agafonkin 2010] Agafonkin, V. (2010). *Leaflet: an open-source JavaScript library for mobile-friendly interactive maps*.

[Aji et al. 2013] Aji, A., Wang, F., Vo, H., Lee, R., Liu, Q., Zhang, X., e Saltz, J. (2013). Hadoop gis: A high performance spatial data warehousing system over mapreduce. *Proc. VLDB Endow.*, 6(11):1009–1020.

[Balasubramanian e Sugumaran 2012] Balasubramanian, L. e Sugumaran, M. (2012). A state-of-art in r-tree variants for spatial indexing. *International Journal of Computer Applications*, 42(20):35–41.

[Barbosa et al. 2018] Barbosa, H., Barthelemy, M., Ghoshal, G., James, C. R., Lenormand, M., Louail, T., Menezes, R., Ramasco, J. J., Simini, F., e Tomasini, M. (2018). Human mobility: Models and applications. *Physics Reports*, 734:1–74.

[Blondel et al. 2015] Blondel, V. D., Decuyper, A., e Krings, G. (2015). A survey of results on mobile phone datasets analysis. *EPJ data science*, 4(1):10.

[Boeing 2017] Boeing, G. (2017). Osmnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. *Computers, Environment and Urban Systems*, 65:126–139.

[Bokeh Development Team 2019] Bokeh Development Team (2019). *Bokeh: Python library for interactive visualization*.

[Bolstad 2016] Bolstad, P. (2016). *GIS Fundamentals: A First Text on Geographic Information Systems*. Eider Press, 5 edition.

[Casanova et al. 2005] Casanova, M. A., Câmara, G., Davis, C., Vinhas, L., e Queiroz, G. R. (2005). *Banco de dados geográficos*. MundoGEO Curitiba.

[Castro et al. 2012] Castro, P. S., Zhang, D., e Li, S. (2012). Urban traffic modelling and prediction using large scale taxi gps traces. In *International Conference on Pervasive Computing*, pages 57–72. Springer.

[Celes et al. 2017] Celes, C., Silva, F. A., Boukerche, A., d. C. Andrade, R. M., e Loureiro, A. A. F. (2017). Improving vanet simulation with calibrated vehicular mobility traces. *IEEE Transactions on Mobile Computing*, 16(12):3376–3389.

[Chen et al. 2019] Chen, G., Viana, A. C., Fiore, M., e Sarraute, C. (2019). Complete trajectory reconstruction from sparse mobile phone data. *EPJ Data Science*, 8(1):30.

[Chen et al. 2017] Chen, G., Viana, A. C., e Sarraute, C. (2017). Towards an adaptive completion of sparse call detail records for mobility analysis. In *2017 IEEE international conference on pervasive computing and communications workshops (PerCom workshops)*, pages 302–305. IEEE.

[Cho et al. 2011] Cho, E., Myers, S. A., e Leskovec, J. (2011). Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1082–1090.

[Choi et al. 2017] Choi, H. J., Hecht, M., e Smith, R. A. (2017). Investigating the potential impact of social talk on prevention through social networks: The relationships between social talk and refusal self-efficacy and norms. *Prevention Science*, 18(4):459–468.

[Claudia Dolci 2010] Claudia Dolci, Dante Salvini, M. S. R. W. (2010). [Online; accessed 19. Apr. 2020].

[Coimbra et al. 2019] Coimbra, G. T., Capanema, C. G. S., Silva, F. A., e Silva, T. R. B. (2019). Appel: Uma extensao do kepler para enriquecimento de dados geoespaciais. *GEOINFO, 20 Years After!*, page 176.

[Cunha et al. 2016] Cunha, F. D., Silva, F. A., Celes, C., Maia, G., Ruiz, L. B., Andrade, R. M. C., Mini, R. A. F., Boukerche, A., e Loureiro, A. A. F. (2016). Communication analysis of real vehicular calibrated traces. In *2016 IEEE International Conference on Communications (ICC)*, pages 1–6.

[de Melo et al. 2015] de Melo, P. O. V., Viana, A. C., Fiore, M., Jaffrès-Runser, K., Le Mouël, F., Loureiro, A. A., Addepalli, L., e Guangshuo, C. (2015). Recast: Telling apart social and random relationships in dynamic networks. *Performance Evaluation*, 87:19–36.

[De Montjoye et al. 2013] De Montjoye, Y.-A., Hidalgo, C. A., Verleysen, M., e Blondel, V. D. (2013). Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, 3:1376.

[Developers 2020] Developers, P. (2020). PostGIS — Documentation. [Online; accessed 15. Apr. 2020].

[Domingues et al. 2018] Domingues, A. C. S. A., Silva, F. A., e Loureiro, A. A. F. (2018). Space and time matter: An analysis about route selection in mobility traces. In *2018 IEEE Symposium on Computers and Communications (ISCC)*, pages 00958–00963.

[Duckham e Kulik 2005a] Duckham, M. e Kulik, L. (2005a). A formal model of obfuscation and negotiation for location privacy. In *International conference on pervasive computing*, pages 152–170. Springer.

[Duckham e Kulik 2005b] Duckham, M. e Kulik, L. (2005b). Simulation of obfuscation and negotiation for location privacy. In *International conference on spatial information theory*, pages 31–48. Springer.

[Ekman et al. 2008] Ekman, F., Keränen, A., Karvo, J., e Ott, J. (2008). Working day movement model. In *Proceedings of the 1st ACM SIGMOBILE workshop on Mobility models*, pages 33–40.

[Finkel e Bentley 1974] Finkel, R. A. e Bentley, J. L. (1974). Quad trees a data structure for retrieval on composite keys. *Acta informatica*, 4(1):1–9.

[Firestone et al. 2011] Firestone, S. M., Ward, M. P., Christley, R. M., e Dhand, N. K. (2011). The importance of location in contact networks: Describing early epidemic spread using spatial social network analysis. *Preventive Veterinary Medicine*, 102(3):185 – 195. Special Issue: GEOVET 2010.

[Gao 2015] Gao, S. (2015). Spatio-temporal analytics for exploring human mobility patterns and urban dynamics in the mobile age. *Spatial Cognition & Computation*, 15(2):86–114.

[González et al. 2008] González, M. C., Hidalgo, C. A., e Barabási, A.-L. (2008). Understanding individual human mobility patterns. *Nature*, 453(7196):779–782.

[Gu et al. 2016] Gu, Y., Yao, Y., Liu, W., e Song, J. (2016). We know where you are: Home location identification in location-based social networks. In *2016 25th International Conference on Computer Communication and Networks (ICCCN)*, pages 1–9. IEEE.

[Guttman 1984] Guttman, A. (1984). R-trees: A dynamic index structure for spatial searching. In *Proceedings of the 1984 ACM SIGMOD international conference on Management of data*, pages 47–57.

[Hess et al. 2015] Hess, A., Hummel, K. A., Gansterer, W. N., e Haring, G. (2015). Data-driven human mobility modeling: a survey and engineering guidance for mobile networking. *ACM Computing Surveys (CSUR)*, 48(3):1–39.

[Hoteit et al. 2016] Hoteit, S., Chen, G., Viana, A., e Fiore, M. (2016). Filling the gaps: On the completion of sparse call detail records for mobility analysis. In *Proceedings of the Eleventh ACM Workshop on Challenged Networks*, pages 45–50. ACM.

[Hung et al. 2009] Hung, C.-C., Chang, C.-W., e Peng, W.-C. (2009). Mining trajectory profiles for discovering user communities. In *Proceedings of the 2009 International Workshop on Location Based Social Networks*, pages 1–8.

[Ingole e Nichat 2013] Ingole, P. e Nichat, M. M. K. (2013). Landmark based shortest path detection by using dijkestra algorithm and haversine formula. *International Journal of Engineering Research and Applications (IJERA)*, 3(3):162–165.

[Johnson e Watson 1984] Johnson, G. T. e Watson, I. D. (1984). The determination of view-factors in urban canyons. *Journal of Climate and Applied Meteorology*, 23(2):329–335.

[Jurdak et al. 2015] Jurdak, R., Zhao, K., Liu, J., AbouJaoude, M., Cameron, M., e Newth, D. (2015). Understanding human mobility from twitter. *PloS one*, 10(7):e0131469–e0131469.

[Kang et al. 2004] Kang, J. H., Welbourne, W., Stewart, B., e Borriello, G. (2004). Extracting places from traces of locations. In *Proceedings of the 2nd ACM international workshop on Wireless mobile applications and services on WLAN hotspots*, pages 110–118. ACM.

[Kolar et al. 2014] Kolar, V., Ranu, S., Subramainan, A. P., Shrinivasan, Y., Telang, A., Kokku, R., e Raghavan, S. (2014). People in motion: Spatio-temporal analytics on call detail records. In *2014 Sixth International Conference on Communication Systems and Networks (COMSNETS)*, pages 1–4. IEEE.

[Kosta et al. 2012] Kosta, S., Mei, A., e Stefa, J. (2012). Large-scale synthetic social mobile networks with swim. *IEEE Transactions on Mobile Computing*, 13(1):116–129.

[Krumm 2009] Krumm, J. (2009). A survey of computational location privacy. *Personal and Ubiquitous Computing*, 13(6):391–399.

[Li et al. 2008] Li, Q., Zheng, Y., Xie, X., Chen, Y., Liu, W., e Ma, W.-Y. (2008). Mining user similarity based on location history. In *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems*, pages 1–10.

[Li et al. 2016] Li, S., Dragicevic, S., Castro, F. A., Sester, M., Winter, S., Coltekin, A., Pettit, C., Jiang, B., Haworth, J., Stein, A., et al. (2016). Geospatial big data handling theory and methods: A review and research challenges. *ISPRS journal of Photogrammetry and Remote Sensing*, 115:119–133.

[Lisboa Filho e Iochpe 2001] Lisboa Filho, J. e Iochpe, C. (2001). Modelagem de bancos de dados geográficos. In *Apostila do XX Congresso Brasileiro de Cartografia, Porto Alegre*.

[Liu et al. 2016] Liu, K., Wang, H., e Yao, Y. (2016). On storing and retrieving geospatial big-data in cloud. In *Proceedings of the Second ACM SIGSPATIALInternational Workshop on the Use of GIS in Emergency Management*, EM-GIS '16, pages 16:1–16:4, New York, NY, USA. ACM.

[Longley et al. 2005] Longley, P. A., Goodchild, M. F., Maguire, D. J., e Rhind, D. W. (2005). *Geographic information systems and science*. John Wiley & Sons.

[Maouche et al. 2017] Maouche, M., Mokhtar, S. B., e Bouchenak, S. (2017). Ap-attack: a novel user re-identification attack on mobility datasets. In *Proceedings of the 14th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, pages 48–57. ACM.

[Marino 2012] Marino, T. (2012). [Online; accessed 15. Apr. 2020].

[Marques-Neto et al. 2018] Marques-Neto, H. T., Xavier, F. H., Xavier, W. Z., Malab, C. H. S., Ziviani, A., Silveira, L. M., e Almeida, J. M. (2018). Understanding human mobility and workload dynamics due to different large-scale events using mobile phone data. *Journal of Network and Systems Management*, 26(4):1079–1100.

[Massey Jr 1951] Massey Jr, F. J. (1951). The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78.

[Montazeri et al. 2017] Montazeri, Z., Houmansadr, A., e Pishro-Nik, H. (2017). Achieving perfect location privacy in wireless devices using anonymization. *IEEE Transactions on Information Forensics and Security*, 12(11):2683–2698.

[Morales et al. 2017] Morales, A. J., Vavilala, V., Benito, R. M., e Bar-Yam, Y. (2017). Global patterns of synchronization in human communications. *Journal of the Royal Society Interface*, 14(128):20161048.

[Morton 1966] Morton, G. M. (1966). A computer oriented geodetic data base and a new technique in file sequencing.

[Motlagh et al. 2016] Motlagh, N. H., Taleb, T., e Arouk, O. (2016). Low-altitude unmanned aerial vehicles-based internet of things services: Comprehensive survey and future perspectives. *IEEE Internet of Things Journal*, 3(6):899–922.

[Naboulsi et al. 2016] Naboulsi, D., Fiore, M., Ribot, S., e Stanica, R. (2016). Large-scale mobile traffic analysis: a survey. *IEEE Communications Surveys & Tutorials*, 18(1):124–161.

[Newson e Krumm 2009] Newson, P. e Krumm, J. (2009). Hidden markov map matching through noise and sparseness. In *Proceedings of the 17th ACM SIGSPATIAL international conference on advances in geographic information systems*, pages 336–343.

[Oracle 2020] Oracle (2020). Spatial and Graph features in Oracle Database | Oracle. [Online; accessed 15. Apr. 2020].

[Praveen et al. 2016] Praveen, P., Babu, C. J., e Rama, B. (2016). Big data environment for geospatial data analysis. In *2016 International Conference on Communication and Electronics Systems (ICCES)*, pages 1–6.

[QGIS Development Team 2009] QGIS Development Team (2009). *QGIS Geographic Information System*. Open Source Geospatial Foundation.

[Rettore et al. 2020] Rettore, P. H., Santos, B. P., Lopes, R. R. F., Maia, G., Villas, L. A., e Loureiro, A. A. (2020). Road data enrichment framework based on heterogeneous data fusion for its. *IEEE Transactions on Intelligent Transportation Systems*, 21(4):1751–1766.

[Sakai et al. 2014] Sakai, T., Tamura, K., e Kitakami, H. (2014). Extracting attractive local-area topics in georeferenced documents using a new density-based spatial clustering algorithm. *IAENG International Journal of Computer Science*, 41(3):185–192.

[Santos et al. 2018] Santos, B. P., Rettore, P. H., Ramos, H. S., Vieira, L. F., e Loureiro, A. A. (2018). Enriching traffic information with a spatiotemporal model based on social media. In *2018 IEEE Symposium on Computers and Communications (ISCC)*, pages 00464–00469. IEEE.

[Shavitt e Zilberman 2011] Shavitt, Y. e Zilberman, N. (2011). A geolocation databases study. *IEEE Journal on Selected Areas in Communications*, 29(10):2044–2056.

[Shimrat 1962] Shimrat, M. (1962). Algorithm 112: position of point relative to polygon. *Communications of the ACM*, 5(8):434.

[Shukla et al. 2016] Shukla, D., Chirag Shivnani, C., e Shah, D. (2016). Comparing oracle spatial and postgres postgis. *IJCSE*, 7:95–100.

[Silva et al. 2015] Silva, F. A., Celes, C., Boukerche, A., Ruiz, L. B., e Loureiro, A. A. (2015). Filling the gaps of vehicular mobility traces. In *Proceedings of the 18th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, pages 47–54.

[Song et al. 2010] Song, C., Qu, Z., Blumm, N., e Barabási, A.-L. (2010). Limits of predictability in human mobility. *Science*, 327(5968):1018–1021.

[Takbiri et al. 2017] Takbiri, N., Houmansadr, A., Goeckel, D. L., e Pishro-Nik, H. (2017). Limits of location privacy under anonymization and obfuscation. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 764–768.

[Teixeira et al. 2019] Teixeira, D. d. C., Viana, A. C., Alvim, M. S., e Almeida, J. M. (2019). Deciphering predictability limits in human mobility. In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 52–61.

[Tong et al. 2011] Tong, X., Wang, Z., Xie, H., Liang, D., Jiang, Z., Li, J., e Li, J. (2011). Designing a two-rank acceptance sampling plan for quality inspection of geospatial data products. *Computers & geosciences*, 37(10):1570–1583.

[Tran et al. 2013] Tran, K. A., Barbeau, S. J., e Labrador, M. A. (2013). Automatic identification of points of interest in global navigation satellite system data: A spatial temporal approach. In *Proceedings of the 4th ACM SIGSPATIAL international workshop on geostreaming*, pages 33–42.

[Uber 2015] Uber (2015). *H3: A hexagonal hierarchical geospatial indexing system*.

[Vis.gl 2018] Vis.gl (2018). *Kepler.gl: a powerful open source geospatial analysis tool for large-scale data sets*.

[Wang e Song 2015] Wang, D. e Song, C. (2015). Impact of human mobility on social networks. *Journal of Communications and Networks*, 17(2):100–109.

[Wang e Taylor 2014] Wang, Q. e Taylor, J. E. (2014). Quantifying human mobility perturbation and resilience in hurricane sandy. *PLoS one*, 9(11).

[Whitman et al. 2014] Whitman, R. T., Park, M. B., Ambrose, S. M., e Hoel, E. G. (2014). Spatial indexing and analytics on hadoop. In *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 73–82. ACM.

[Yu et al. 2015] Yu, J., Wu, J., e Sarwat, M. (2015). Geospark: A cluster computing framework for processing large-scale spatial data. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, page 70. ACM.

[Zhang et al. 2020] Zhang, J., Yang, C., Yang, Q., Lin, Y., e Zhang, Y. (2020). Hgeohashbase: an optimized storage model of spatial objects for location-based services. *Frontiers of Computer Science*, 14(1):208–218.

[Zheng et al. 2014] Zheng, Y., Capra, L., Wolfson, O., e Yang, H. (2014). Urban computing: concepts, methodologies, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(3):1–55.