Capítulo 2

Computação Urbana: Técnicas para o Estudo de Sociedades com Redes de Sensoriamento Participativo

Thiago H. Silva e Antonio A. F. Loureiro

Abstract

Urban computing is a recent research topic that aims to obtain and analyze urban data from various sources, such as traditional wireless sensor networks (WSNs) and emerging participatory sensor networks (PSNs) to understand and address issues that cities face. PSNs are particularly interesting because they rely on the participation of users in urban sensing, allowing the observation of large-scale actions of people in (almost) real time over long periods of time. PSN data increase our knowledge about different aspects of our lives in urban scenarios, which can be very useful in developing more sophisticated applications to various sectors, especially those related to the understanding of urban societies. The purpose of this short course is to discuss the concept of urban computing and urban sensing with participatory sensor networks. We aim to show the relevance of urban computing and motivate the construction of new applications that address issues related to the dynamics of cities and urban social behavior. In addition, this short course will discuss how to work with PSNs, by analyzing their properties and their usefulness in the development of new applications in urban computing.

Resumo

A computação urbana (urban computing) é um tema recente de pesquisa que visa obter e analisar dados urbanos de diversas fontes, como as tradicionais redes de sensores sem fio (RSSFs) e as emergentes redes de sensoriamento participativo (RSP), com o objetivo de entender e tratar questões enfrentadas pelas cidades. As RSPs são particularmente interessantes nesse caso, pois contam com a colaboração dos usuários no sensoriamento urbano, permitindo a observação das ações das pessoas em larga escala em tempo (quase) real durante longos períodos de tempo. Dados de RSPs aumentam o nosso conhecimento sobre diferentes aspectos de nossas vidas em cenários urbanos o que pode ser bastante útil no desenvolvimento de aplicações mais sofisticadas para diversos segmentos, principalmente os relacionados com o entendimento de sociedades urbanas. O objetivo deste minicurso é discutir o conceito de computação urbana e de sensoriamento urbano com redes de sensoriamento participativo. Visamos mostrar a relevância da computação urbana e motivar a

construção de novas aplicações que sirvam para tratar questões relacionadas com a dinâmica de cidades e do comportamento social urbano. Além disso, este minicurso discutirá como trabalhar com RSPs, ao analisar as suas propriedades e a sua utilidade no desenvolvimento de novas aplicações na área de computação urbana.

2.1. Introdução

A computação urbana (*urban computing*) [Kindberg et al. 2007, Kostakos and O'Neill 2008, Zheng et al. 2014a] é uma área interdisciplinar que diz respeito ao estudo e tratamento de questões enfrentadas pelas cidades utilizando tecnologia de computação. Por essa razão, a computação urbana conta com profissionais e aplicações em campos que incluem: antropologia, planejamento urbano, engenharia civil, ciência da computação, entre outros.

Como mais de 50% da população do mundo hoje vive em cidades [Martine et al. 2007], uma das consequências é uma enorme pressão sobre as suas infraestruturas, como transporte, habitação, água e energia, gerando difíceis desafios. Para entender e tratar essas e outras questões com o intuito de melhorar a qualidade de vida das pessoas que vivem em cidades, na computação urbana são usadas diversas fontes de dados sobre o ambiente urbano, alguns exemplos incluem: as tradicionais redes de sensores sem fio (RSSFs) [Loureiro et al. 2003]; e as emergentes redes de sensoriamento participativo (RSP) [Burke et al. 2006, Silva et al. 2014a].

As RSPs são particularmente interessantes nesse caso, pois contam com a colaboração dos usuários no sensoriamento urbano e permitem a observação em larga escala das ações das pessoas em tempo (quase) real durante longos períodos de tempo, possibilitando o entendimento da dinâmica da cidade e do comportamento social urbano. Com isso, as RSPs têm o potencial de se tornarem ferramentas fundamentais para a computação urbana. Dados de RSPs aumentam o nosso conhecimento sobre diferentes aspectos de nossas vidas em cenários urbanos o que pode ser bastante útil no desenvolvimento de aplicações mais sofisticadas em diversos segmentos, principalmente os relacionados com o entendimento de sociedades urbanas.

O objetivo deste minicurso é apresentar o conceito de computação urbana e de sensoriamento urbano com o auxílio de redes de sensoriamento participativo. Isso inclui uma visão geral de trabalhos específicos que ilustram as tendências de pesquisa e os principais desafios e oportunidades da área.

O restante do capítulo está organizado da seguinte forma. A Seção 2.2 apresenta o conceito de computação urbana, incluindo um arcabouço para o desenvolvimento de aplicações nessa área e algumas das principais fontes de dados. A Seção 2.3 discute em mais detalhes uma das fontes de dados urbanos: as redes de sensoriamento participativo. A Seção 2.4 discute o gerenciamento de dados urbanos, o que inclui a obtenção e tratamento desses dados. A Seção 2.5 analisa dados urbanos de RSPs, apresentando algumas de suas principais propriedades. A Seção 2.6 apresenta as abordagens e modelos utilizados em diversas aplicações e serviços relacionados ao estudo de sociedades urbanas utilizando dados de RSPs. A Seção 2.7 discute as principais técnicas utilizadas nos trabalhos mencionados nas seções anteriores, bem como algumas das tecnologias e ferramentas comumente utilizadas para a análise de dados. A Seção 2.8 apresenta alguns dos principais desafios relacionados

com a utilização de RSPs na computação urbana, já a Seção 2.9 apresenta várias oportunidades nessa mesma direção. Finalmente, a Seção 2.10 apresenta as nossas conclusões.

2.2. Computação Urbana

2.2.1. Definição

O termo "computação urbana" foi introduzido pela primeira vez por Eric Paulos na edição de 2004 da conferência UbiComp [Eric Paulos and Townsend 2004] e em seu artigo *The Familiar Stranger* [Paulos and Goodman 2004], publicado nesse mesmo ano.

Pode-se definir a computação urbana como um processo de aquisição, integração e análise de um grande volume de dados heterogêneos gerados por diversas fontes em espaços urbanos, tais como sensores, veículos e seres humanos, para ajudar na solução de diversos problemas que as cidades enfrentam tais como congestionamento de trânsito, poluição do ar, falta de água e aumento do consumo de energia. Assim um dos principais objetivos dessa área é ajudar a melhorar a qualidade de vida das pessoas que vivem em ambientes urbanos [Zheng et al. 2014a].

A computação urbana também nos auxilia a compreender a natureza dos fenômenos urbanos, bem como prever o futuro das cidades. Essa é uma área bastante interdisciplinar resultante da fusão da área de ciência da computação com áreas tradicionais, como transporte, economia e sociologia no contexto dos espaços urbanos. No domínio da ciência da computação, a computação urbana tem interseção com, por exemplo, sistemas distribuídos, interação humano-computador, redes de computadores, redes de sensores, sistemas cooperativos e inteligência artificial.

2.2.2. Arcabouço da Computação Urbana

Nesta seção apresentamos um arcabouço para a computação urbana. A Figura 2.1 mostra uma visão geral desse arcabouço, destacando os três componentes mais importantes: (i) gerenciamento dos dados urbanos; (ii) análise dos dados urbanos; e (iii) desenvolvimento de serviços e aplicações.

Figura 2.1. Visão geral do arcabouço da computação urbana.

Como ilustrado na figura, o componente gerenciamento de dados urbanos é composto de alguns passos importantes. O primeiro deles é o processo de coleta de dados urbanos, que podem ser obtidos de diversas fontes de dados, como discutido na próxima se-

ção. O segundo passo refere-se ao processamento desses dados. Após esse processamento podemos modelar os dados em diversos formatos, por exemplo, no formato de grafos, como discutido na Seção 2.4.

O componente análise dos dados urbanos é composto pela etapa de edição e execução de códigos, bem como a interpretação de resultados. Essa parte é fundamental, pois para utilizar dados urbanos é necessário conhecer suas propriedades. Mais detalhes sobre esse componente é descrito na Seção 2.5. Após a etapa de análise, o próximo passo é o desenvolvimento de serviços e aplicações com o conhecimento obtido. Essas aplicações podem ser de diversos tipos, como discutido na Seção 2.6.

2.2.3. Fontes de Dados Urbanos

Nesta seção apresentamos algumas das principais fontes de obtenção de dados urbanos. Esses dados oferecem suporte no desenvolvimento de novos serviços e aplicações na área de computação urbana.

- **Dados estatísticos oficiais:** fornecem dados referentes a um estudo estatístico sobre uma população, tais como dados demográficos, econômicos e sociais relativos a um momento determinado ou em certos períodos.

  É possível encontrar diversas fontes de dados na Web disponibilizando dados dessa categoria para algumas localidades, como mostrado em [Barbosa et al. 2014]. No entanto, nem sempre esses dados estão disponíveis para a localidade que se deseja estudar. Outra dificuldade é a diversidade dos formatos nos quais os dados estão disponíveis, como em tabelas, mapas, gráficos, calendários, formulários, entre outros [Barbosa et al. 2014].

- **Redes de sensores tradicionais:** fornecem dados que são obtidos através da instalação de sensores específicos para algumas aplicações, por exemplo, sensores de presença em ruas e avenidas para detectar o volume de tráfego nesses locais, sensores para monitoramento da qualidade do ar em diversos pontos da cidade ou sensores para o monitoramento de níveis de ruídos.

  Um problema com essa fonte de dados é a dificuldade de acesso aos dados. Além do custo de construção de uma rede de sensoriamento, geralmente, a implantação de sensores na cidade só é permitida pela prefeitura.

- **Infraestrutura das cidades:** fornecem dados que são capturados aproveitando as infraestruturas existentes da cidade, que são criadas para outros propósitos. Por exemplo, as redes de telefonia celular são construídas para comunicação móvel entre os indivíduos. No entanto, os sinais dos telefones celulares de um grande número de pessoas podem ser usados para tentar predizer a mobilidade dos usuários e melhorar o planejamento urbano.

  Outros tipos incluem a localização de veículos que possuem GPS. É cada vez mais comum ônibus, táxis, e veículos privados possuírem GPS embutidos. Esse tipo de dado contribui, por exemplo, para o entendimento do tráfego de uma cidade. Além disso, é possível obter dados de utilização do sistema de transporte público, já que é

bem comum esse tipo de sistema utilizar cartões RFID para registrar o uso de ônibus e metrô dos usuários.

A dificuldade de acesso a dados é também um problema dessa fonte, uma vez que somente a prefeitura ou empresas responsáveis, tipicamente, possuem acesso a esse tipo de dado.

- **Redes de sensoriamento participativos:** fornecem dados urbanos, que possuem uma escala bastante abrangente e podem ser mais fáceis de obter do que as outras fontes mencionadas, pois contam com a colaboração dos usuários na coleta de dados. Além disso, as RSPs podem possuir uma rede social online o que permite o estudo da estrutura social dos usuários, como relacionamentos e interações entre os usuários.

2.3. Sensoriamento Urbano com Redes de Sensoriamento Participativo

Como apresentamos na Seção 2.2, existem várias formas de obter dados urbanos, dentre elas podemos citar as emergentes redes de sensoriamento participativo (RSPs) [Silva et al. 2014a, Burke et al. 2006]. O tema será abordado da seguinte maneira: a Seção 2.3.1 apresenta a definição de uma RSP; a Seção 2.3.2 discute o funcionamento de uma RSP, enquanto a Seção 2.3.3 ilustra exemplos de RSPs.

2.3.1. O que é uma rede de sensoriamento participativo?

O sensoriamento participativo pode ser definido como um processo distribuído de coleta de dados pessoais e sobre diversos aspectos da cidade. Tal processo requer a participação ativa das pessoas para compartilhar voluntariamente informação contextual e/ou tornar seus dados sensoriados disponíveis [Burke et al. 2006], ou seja, o usuário determina manualmente como, quando, o quê e onde amostrar. Assim, através das RSPs é possível monitorar diversos aspectos das cidades, bem como o comportamento coletivo de pessoas conectadas à Internet em tempo (quase) real.

As RSPs têm se tornado populares graças ao aumento do uso de dispositivos portáteis, como *smartphones* e *tablets*, assim como a adoção mundial de sites de mídia social. Com isso, um elemento central de uma rede de sensoriamento participativo é a existência de um usuário capaz de realizar um sensoriamento, por exemplo, da cidade, com um dispositivo computacional portátil. Nesse cenário, as pessoas participam como sensores sociais, fornecendo dados voluntariamente sobre um determinado aspecto de um local que implicitamente capturam as suas experiências de vida diária. Esses dados podem ser obtidos com a ajuda de dispositivos de sensoriamento, como sensores incorporados a *smartphones* (e.g., GPS, acelerômetro, microfone, e outros), ou por meio de sensores humanos (e.g., visão). Nesse último caso, os dados são observações subjetivas produzidas pelos usuários [Silva et al. 2014a, Burke et al. 2006].

As RSPs oferecem oportunidades sem precedentes de acesso a dados de sensoriamento em escala planetária. Essa grande quantidade de dados facilita a obtenção de informações que não estão disponíveis prontamente com a mesma abrangência global, podendo ser usadas para melhorar os processos de tomada de decisão de diferentes entidades (e.g., pessoas, grupos, serviços, aplicações).

Vale ressaltar que vários termos definidos recentemente como, por exemplo, *Humans as Data Sources* e *Ubiquitous Crowdsourcing*, refletem basicamente a definição de redes de sensoriamento participativo considerada neste documento [Srivastava et al. 2012, Mashhadi and Capra 2011, Ganti et al. 2011]. É importante também mencionar que o termo sensoriamento oportunista [Lane et al. 2010], que denomina uma forma de sensoriamento que também utiliza dispositivos móveis dos usuários no processo de sensoriamento, pode gerar confusão com o termo sensoriamento participativo. O sensoriamento participativo difere de sensoriamento oportunista principalmente pela participação do usuário, onde, neste último tipo, a etapa de coleta de dados é automatizada, sem a participação do usuário [Lane et al. 2008, Lane et al. 2010].

O sensoriamento oportunista apoia o processo de sensoriamento de uma aplicação sem demandar esforços do usuário, determinando automaticamente quando os dispositivos podem ser usados para atender às demandas específicas das aplicações. Desta forma, os aplicativos podem aproveitar as capacidades de sensoriamento de todos os dispositivos dos usuários do sistema sem a necessidade de intervenção humana neste processo [Lane et al. 2008].

2.3.2. O funcionamento de uma RSP

De forma similar às tradicionais redes de sensores sem fio (RSSFs) [Loureiro et al. 2003], o dado sensoriado em uma RSP é enviado para o servidor, ou "nó sorvedouro", onde os dados podem ser acessados (usando, por exemplo, APIs, como a API do Instagram[1]). Mas, diferentemente das RSSFs, as RSPs têm as seguintes características: (a) nós sensores são entidades móveis autônomas, ou seja, uma pessoa com um dispositivo móvel; (b) o custo da rede é distribuído entre os nós sensores, proporcionando uma escalabilidade global; (c) o sensoriamento depende da vontade das pessoas participarem desse processo; e (d) nós sensores não possuem severas limitações de energia.

As RSPs têm o potencial para complementar as RSSFs em diversos aspectos. As tradicionais redes de sensores sem fio foram projetadas para sensoriar áreas de tamanho limitado, como florestas e vulcões. Em contrapartida, as RSPs podem alcançar áreas de tamanhos variados e de larga escala, como grandes metrópoles, países ou até mesmo todo o planeta [Silva et al. 2014a]. Além disso, uma RSSF está sujeita a falhas, uma vez que o seu funcionamento depende da correta coordenação das ações dos seus nós sensores que possuem severas restrições de energia, processamento e memória. Já as RSPs são formadas por entidades autônomas e independentes, os seres humanos, o que torna a tarefa de sensoriamento mais robusto a falhas individuais. Obviamente, RSPs trazem também vários novos desafios como, por exemplo, o seu sucesso está diretamente ligado à popularização dos *smartphones*, *tablets* e serviços que operam na Internet, principalmente as mídias sociais.

A Figura 2.2 ilustra uma RSP constituída de usuários com seus dispositivos móveis enviando dados sensoriados sobre suas localizações para sistemas na nuvem. A figura mostra as atividades de compartilhamento (representados por pontos na nuvem) de quatro usuários em três instantes diferentes no tempo, rotulados como "Tempo 1", "Tempo 2" e "Tempo 3". Note que um usuário não participa, necessariamente, no sistema em todos os instantes. Após um certo tempo, podemos analisar estes dados de diferentes maneiras. Por exemplo, a parte inferior mais à direita da figura mostra, por meio de

[1] http://instagram.com/developer.

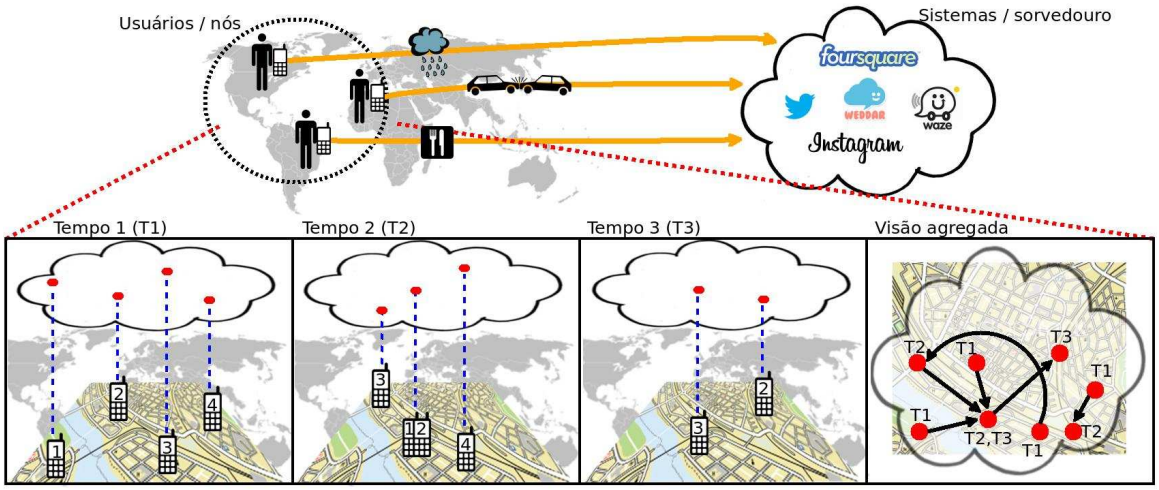


Figura 2.2. Ilustração de uma rede de sensoriamento participativo [Silva et al. 2014a].

uma visão agregada, um grafo dirigido em que os nós/vértices representam os locais onde os dados foram compartilhados e com arestas que conectam localidades que foram compartilhadas pelo mesmo usuário. Usando este grafo podemos extrair, por exemplo, padrões de mobilidade dos usuários que podem ser utilizados para efetuar um gerenciamento de carga de forma mais eficiente na infraestrutura urbana de redes sem fio. Na verdade, a descoberta de conhecimento em RSPs caminha junto com o uso da teoria de grafos/redes [Easley and Kleinberg 2010, Newman 2010, Newman 2003].

2.3.3. Exemplos de RSPs

As redes sociais baseadas em localização, que são um tipo especial de mídia social que combinam características de rede social *online*[2] e a possibilidade de compartilhar dados com informações espaço-temporais[3], podem ser consideradas os exemplos mais populares de RSPs. É possível encontrar vários exemplos de tais sistemas em funcionamento, tais como: Waze, que serve para relatar condições de tráfego em tempo real; Foursquare, para compartilhar o local onde o usuário está visitando; e Instagram, para enviar imagens em tempo real para o sistema. Em particular, o Instagram pode ser visto como uma das mais populares RSPs atualmente, com 200 milhões de usuários [Instagram 2014]. Ao considerarmos essa rede, o dado sensoriado é uma foto de um lugar específico. Podemos extrair informação desse tipo de dado de diversas maneiras. Uma das possibilidades é visualizar em tempo real como está a situação de uma certa área da cidade. Outras possibilidades são discutidas na Seção 2.6.

Note que todos os sistemas descritos anteriormente são compostos de uma rede social *online*. No entanto, existem vários exemplos de RSPs que não contêm redes sociais *online*. Por exemplo, o Weddar[4], para relatar condições meteorológicas, o NoiseTube[5], para

[2]Plataforma virtual que constrói e reflete as relações sociais da vida real entre as pessoas.
[3]Tipo de dado que permite, por exemplo, a construção de serviços baseados em localização.
[4]http://www.weddar.com.
[5]http://noisetube.net.

o compartilhamento de nível de ruído em determinada região da cidade ou o Colab[6], para o compartilhamento de problemas diversos das cidades.

Alguns outros tipos de mídia social, como o Twitter[7], que permite aos seus usuários compartilhar atualizações pessoais em textos de até 140 caracteres, conhecidos como *tweets*, podem também ser exemplos de RSPs. Twitter é considerado um exemplo de RSP porque *tweets* podem, eventualmente, também permitir a monitorização de vários aspectos das cidades, bem como o comportamento coletivo das pessoas quase em tempo real. Por exemplo, as pessoas poderiam usar os seus dispositivos portáteis para compartilhar *tweets* com informações em tempo real sobre manifestações ou acidentes na cidade. Além desses exemplos, podemos também mencionar GarbageWatch [CENS/UCLA] para monitorar aspectos de lixo de uma cidade. Este exemplo é particularmente interessante porque ilustra que a utilização da web não é obrigatória em uma RSP. Dado de sensoriamento pode ser enviado para um aplicativo específico em execução na Internet, mas fora da Web.

2.4. Gerência de Dados Urbanos

2.4.1. Obtenção de Dados

Nesta seção apresentamos três das principais formas de obtenção de dados de RSPs. A primeira forma é através de APIs, como descrita na Seção 2.4.1.1. A segunda forma é utilizando um Web *crawler* (Seção 2.4.1.2). Por fim, apresentamos na Seção 2.4.1.3 a forma de obtenção de dados que utiliza aplicações.

2.4.1.1. Utilizando APIs

A Web está repleta de fontes de informação, o que representa uma grande oportunidade para pesquisadores de diversas áreas coletarem dados em larga escala e a partir deles extrair conhecimento [Benevenuto et al. 2011].

Algumas RSPs disponibilizam APIs que podem ser utilizadas para a extração de dados. Através desse processo, é possível obter dados de RSPs que podem ser utilizados em outras aplicações ou em análises específicas. Várias RSPs populares, como Foursquare, possuem APIs de acesso aos dados compartilhados pelos usuários. Entretanto, é comum existirem regras diferentes para a sua utilização.

Podemos citar duas formas de funcionamento de APIs: (1) baseadas em *streaming*; (2) baseadas em requisições. O método baseado em *streaming* permite coletar em tempo (quase) real os dados que são publicados em uma determinada RSP. A API de *streaming* do Twitter[8], por exemplo, permite coletar em tempo (quase) real *tweets* públicos à medida que são publicados. Já o método baseado em requisições disponibilizam dados atendendo a uma solicitação específica, por exemplo, todos os seus últimos 10 *tweets*. Tanto métodos baseados em *streaming* quanto métodos baseados em requisições podem sofrer limitações na obtenção do volume de dados. Por exemplo, o Flickr permite 5000 requisições por hora em sua API, já a API de *streaming* do Twitter pode não fornecer todos os dados comparti-

[6]www.colab.re.
[7]http://www.twitter.com.
[8]http://www.twitter.com.

lhados[9]. Isso pode inviabilizar alguns tipos de análises que necessitam de um número maior de amostras no período de uma hora, por exemplo.

De fato, o uso de APIs é uma forma bastante popular para a obtenção de dados. Dados obtidos através de APIs como a do Twitter foram utilizados das mais variadas formas, desde medir a influência de usuários na rede [Cha et al. 2010], até a previsão de terremotos [Sakaki et al. 2010a].

Um exemplo de uso da API de *streaming* do Twitter, escrito na linguagem de programação Python e utilizando a biblioteca TwitterAPI[10], é mostrado no algoritmo mostrado na Figura 2.3. Nesse algoritmo fazemos acesso aos *tweets* buscando pela palavra-chave "4sq". Como podemos ver, em poucas linhas de código é possível coletar facilmente dados do Twitter. A Figura 2.4 ilustra esse resultado com dois *tweets* de resposta: tweet1 e tweet2. Esses *tweets* foram retornados no padrão JSON.

```
#Biblioteca que facilita a interação com a API DO Twitter
from TwitterAPI import TwitterAPI

#Um registro no website da API fornece as credenciais indicadas aqui
twitter_api = TwitterAPI(consumer_key='xxxxxx',  consumer_secret='xxxxxx',
»       access_token_key='xxxxxxx', access_token_secret='xxxxxxx')

filters = {"track": ["4sq"]} #palavra que deseja buscar em tweets

stream = twitter_api.request('statuses/filter', filters)

for item in stream.get_iterator():
  print item #exibe todo conteúdo do tweet
```

Figura 2.3. Exemplo de obtenção de dados do Twitter.

Existem RSPs que possuem APIs, mas com acesso bastante restrito aos dados. Esse é o caso do Foursquare, pois poucos dados são possíveis de serem coletados sem a autorização do usuário. A maioria dos dados disponíveis através dessa API são referentes aos locais, como dicas, listas, localização e fotos.

Essas limitações estimulam a obtenção de dados de forma indireta ou alternativa. Por exemplo, em [Silva et al. 2014c] os autores obtiveram dados sobre os *check-ins* do Foursquare através de mensagens públicas compartilhadas no Twitter. Isso é possível, pois o Foursquare possibilita aos usuários anunciarem seus *check-ins* nesse sistema. Esse procedimento é mostrado na Figura 2.4. Essa figura ilustra um *tweet* proveniente do Foursquare que possui uma URL que representa uma página Web com mais informações sobre o *check-in* anunciado. No exemplo da figura, a página representa um *check-in* dado em um restaurante. Para obter mais dados sobre o *check-in* disponíveis nessa página é utilizada a técnica de coleta Web *crawler*, apresentada a seguir.

2.4.1.2. Utilizando um Web *Crawler*

Nem todas as fontes de dados disponíveis na Internet fornecem acesso direto a esses dados através de APIs. Por isso é necessário utilizar outras formas de obtenção de dados. Uma

[9]Ao solicitar dados os dados compartilhados no Twitter estima-se que será entregue 1%.
[10]https://github.com/geduldig/TwitterAPI.

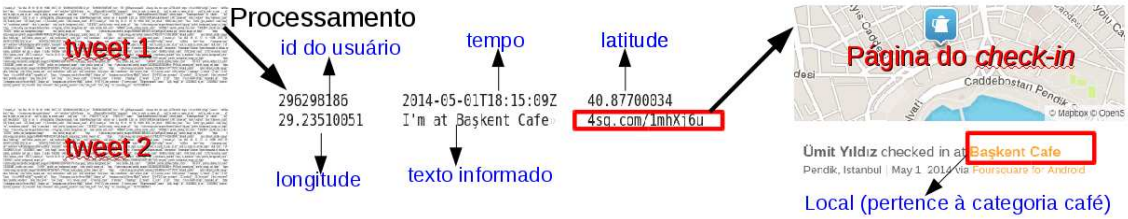


Figura 2.4. Etapas de coleta de dados do Foursquare através de *tweets*.

dessas alternativas é a chamada Web *crawler*, que são programas que analisam páginas Web em busca de dados relevantes [Benevenuto et al. 2011]. Um Web *crawler* funciona como um robô que acessa páginas Web predeterminadas e recupera dados a partir dessas páginas.

A coleta através de Web *crawlers* depende da estrutura da fonte da qual desejamos obter dados, bem como da abordagem utilizada. A estrutura da fonte é onde os dados que queremos extrair estão disponibilizados; nas páginas Web, por exemplo, são *tags* HTML que apresentam os dados ao usuário. Com isso a construção de um Web *crawler* demanda tipicamente a mineração de texto para a extração dos dados necessários na página Web estudada. No entanto, outras formas não convencionais de extração de dados usando páginas Web são possíveis. Por exemplo, em [Tostes et al. 2014] os autores construíram um Web *crawler* para coletar informações de tráfego tirando fotos (*screenshots*) de mapas com essas informações, como as disponíveis no Bing Maps[11]. Mais informações sobre esse procedimento são fornecidas em [Tostes et al. 2014].

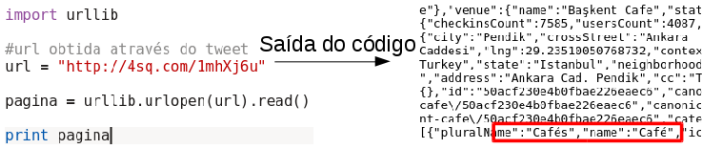


Figura 2.5. Exemplo de coleta de uma página de um *check-in* do Foursquare usando a biblioteca URLLIB da linguagem Python.

A Figura 2.5 ilustra um código de um simples Web *crawler* em Python. Esse Web *crawler* utiliza a biblioteca URLLIB para realizar a coleta de uma página referente a um check-in do Fousquare (a URL utilizada foi a encontrada no processo ilustrado na Figura 2.4). O resultado parcial da saída do código é também ilustrado nessa figura. Repare que foi assinalada uma informação relevante sobre o *check-in*: o tipo do local, informação que não é acessível através da API do Foursquare.

2.4.1.3. Utilizando Aplicações

Uma outra alternativa para a coleta de dados é a criação de aplicações em plataformas já existentes. Alguns sistemas populares, como Facebook, Instagram e Runkeeper, permitem a criação de aplicativos dentro de suas plataformas. Com isso, desenvolvedores podem oferecer serviços utilizando dados que são compartilhados nesses aplicativos.

[11] http://www.bing.com/maps.

O Facebook, por exemplo, não permite a coleta de informações direta de seus usuários por APIs ou Web *crawlers*. No entanto, como permitem a criação de aplicações é possível obter dados compartilhados por seus usuários. Quando o usuário do Facebook instala um aplicativo e autoriza a leitura de seus dados, o desenvolvedor da aplicação pode ler e armazenar diversos dados, como os disponibilizados pelos usuários, por exemplo, o conteúdo compartilhado com seus amigos.

Em [Nazir et al. 2008] os autores utilizaram essa abordagem de coleta de dados. Eles criaram aplicações do Facebook especificamente para coletar dados que possibilitassem o estudo do comportamento das pessoas que fazem uso desse tipo de aplicação. Outro exemplo foi o aplicativo utilizado em [Youyou et al. 2015]. Os autores criaram uma aplicação no Facebook que captura os últimos *likes*[12] do usuário para traçar um perfil de personalidade.

É possível ainda a criação de aplicações que não dependem de plataformas de sistemas existentes. Esse foi o caso da RSP NoiseTube [Maisonneuve et al. 2009]. Os autores criaram uma aplicação que permite aos usuários reportarem níveis de ruído na cidade. Esses dados permitem identificar, por exemplo, quais áreas da cidade o nível de ruído está acima dos limites estipulados por lei. Outro exemplo é o Colab, citado anteriormente. Recentemente foi proposta uma plataforma chamada *ohmage*[13] para facilitar a construção de aplicações que desejam utilizar dados de sensoriamento participativo.

Dessa forma, de posse de dados de RSPs, que podem ser obtidos por alguma dessas maneiras citadas, podemos extrair conhecimento de diversas formas, como é melhor discutido nas próximas seções

2.4.2. Reformatação e Limpeza dos Dados

Os dados brutos (sem tratamento) de RSPs podem não são estar em um formato conveniente para executar uma análise particular. Dependendo do tipo do dado é possível encontrar erros semânticos, entradas ausentes ou formatação inconsistente. Nestes casos, eles precisam ser "limpos" antes da análise.

Programadores reformatam e limpam dados escrevendo *scripts* ou editando manualmente dados, por exemplo, numa planilha. Estas tarefas tendem a ser demoradas e tediosas, pois são tarefas inevitáveis que não produzem novos conhecimentos. No entanto, a tarefa de reformatação de dados e de limpeza pode proporcionar ideias sobre quais suposições são seguras de serem feitas sobre os dados, quais peculiaridades existentes no processo de coleta e quais os modelos e análises são apropriados para serem aplicados.

A integração de dados é um desafio relacionado nesta fase, mas é discutido na Seção 2.8. Muitas vezes, a programação envolve trabalhar os dados em diferentes ferramentas, convertendo de um formato de dados para outro, extraindo dados numéricos a partir de um texto, e administrando experimentos numéricos que envolvem um grande número de arquivos de dados e diretórios. Tais tarefas são muito mais rápidas para serem realizadas em uma linguagem como Python do que em Java ou C++, discutimos mais sobre isso na Seção 2.7.10.

[12]Um *like* é uma interação do usuário com o Facebook em que ele demonstra que gostou de um item compartilhado.

[13]http://ohmage.org.

2.4.3. Modelagem de Dados

Os dados gerados em espaços urbanos são geralmente associados com uma propriedade espacial ou espaço-temporais. Por exemplo, a localização de estabelecimentos são dados espaciais; dados meteorológicos e consumo de energia são dados temporais (também chamados de séries temporais, ou *stream*). Já os dados de RSPs possuem propriedades espaço-temporais simultaneamente.

Existem várias formatos de dados para modelar dados de RSPs, sendo bastante popular o uso de grafos [Zheng et al. 2014a]. Na Figura 2.2 ilustramos a criação de um grafo dirigido em que os nós representam os locais onde os dados foram compartilhados e com arestas que conectam localidades que foram compartilhadas pelo mesmo usuário. Usando este grafo podemos extrair diversas informações. De fato, a descoberta de conhecimento em dados de RSPs caminha junto com uma vasta gama de estudos que utilizam a teoria de grafos [Newman 2010, Newman 2003, Easley and Kleinberg 2010]. Como mostramos na Seção 2.6, técnicas bem conhecidas utilizadas para análise de grafos podem ser aplicadas diretamente para estudar grafos derivados de RSPs que refletem condições das cidades.

Alguns dos desafios sobre questões da dinâmica temporal relacionados com dados urbanos das RSPs são discutidos na Seção 2.8.

2.5. Análise de Dados Urbanos Provenientes de RSPs

Como os dados urbanos provenientes de RSPs podem ser muito complexos, um passo fundamental em qualquer investigação é caracterizar os dados coletados, a fim de entender suas limitações e utilidade. Com isso, nesta seção vamos estudar as propriedades de três RSPs para compartilhamento de localização (Foursquare, Gowalla e Brightkite[14]); uma RSP para compartilhamento de fotos (Instagram); bem como uma RSP para compartilhamento de alerta de trânsito (Waze).

2.5.1. Descrição dos Dados

Nesta subseção apresentaremos todos os *datasets* aqui considerados. Todos os dados foram coletados através do Twitter. Além de *tweets* de texto simples, os usuários também podem anunciar dados a partir de uma integração com outros serviços, como o Instagram, Foursquare e Waze. Neste caso, fotos do Instagram, *check-ins* do Foursquare ou alertas do Waze anunciadas no Twitter passam a ficar disponíveis publicamente, o que por padrão não acontece quando o dado é publicado unicamente nos sistemas analisados.

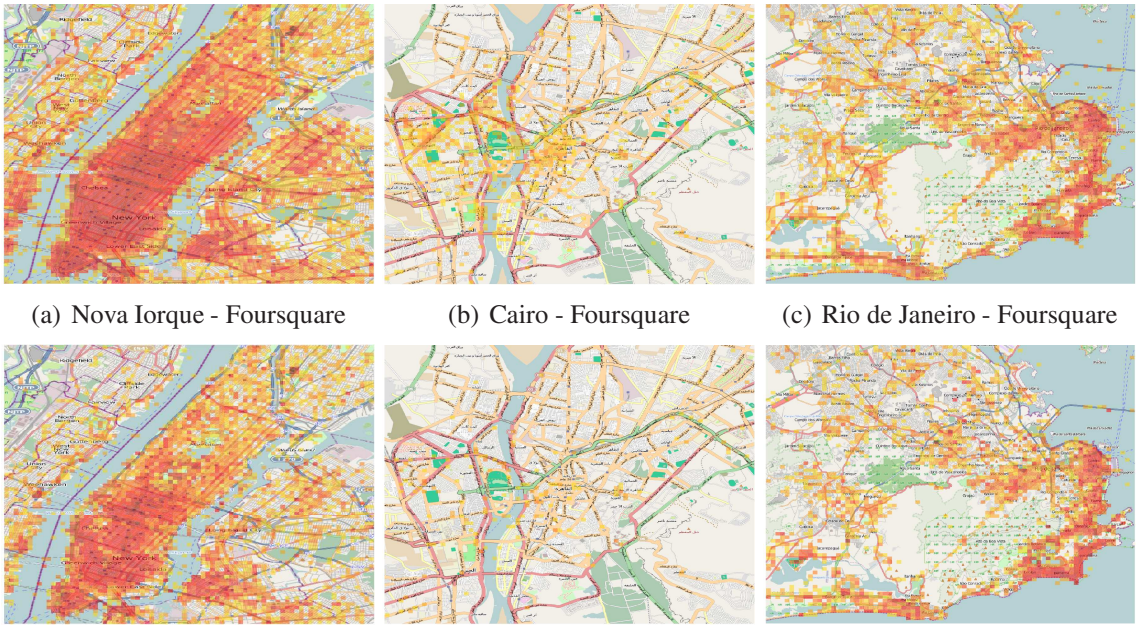
Alguns dos *datasets* que foram analisados: Foursquare1 (≈5 milhões de check-ins em abril de 2012 - 1 semana); Foursquare2 (≈12 milhões de check-ins entre fev2010-jan2011); Foursquare3 (≈4 milhões de check-ins em maio de 2013 - 2 semanas; Gowalla (≈6 milhões de check-ins entre fev2009-out2010); Brightkite (≈4 milhões check-ins entre abr2008-out2010); Instagram1 (≈2 milhões de fotos entre jun2012-jul2012); Instagram2 (≈2 milhões de fotos em maio 2013 - 2 semanas); Waze (+212 mil alertas entre dez2012-jun2013). Como podemos ver, os dados refletem diferentes períodos. Além disso, os *datasets* incluem uma quantia bastante significativa de dados: mais de 30 milhões de registros considerando todas as fontes.

[14] As RSPs para compartilhamento de localização Gowalla e Brightkite não estão mais em funcionamento.

Cada dado sensoriado (foto, *check-in* ou alerta) é composto de coordenadas GPS (latitude e longitude), do horário do compartilhamento do dado e do ID do usuário compartilhador. O *dataset* Foursquare1 possui informações extras sobre o tipo de local: categoria (por exemplo, comida) e um identificador do local. Mais informações sobre os *datasets* e como eles foram obtidos podem ser encontradas em [Cheng et al. 2011, Silva et al. 2012, Silva et al. 2013c, Silva et al. 2013d, Silva et al. 2013e].

2.5.2. Cobertura da Rede

Nesta seção, estudamos a cobertura das RSPs analisadas em diferentes granularidades espaciais, começando por todo o planeta, depois cidades e, por fim, áreas específicas de uma cidade. A primeira constatação ao analisar esses dados é que a cobertura é bastante abrangente e tem escala planetária [Cheng et al. 2011, Silva et al. 2013a, Silva et al. 2013e].



(a) Nova Iorque - Foursquare (b) Cairo - Foursquare (c) Rio de Janeiro - Foursquare

(d) Nova Iorque - Instagram (e) Cairo - Instagram (f) Rio de Janeiro - Instagram

Figura 2.6. Cobertura espacial da RSP do Foursquare e Instagram em 3 cidades populosas ao redor do mundo [Silva et al. 2013a, Silva et al. 2013c].

Avaliamos agora a participação dos usuários em diversas cidades grandes, localizadas em regiões distintas, mostrando os resultados de algumas delas: Nova York, Rio de Janeiro e Cairo. A figura 2.6 mostra o mapa de calor da atividade de sensoriamento para cada uma dessas cidades. Mais uma vez, cores mais escuras representam um maior número de fotos em determinada área. Observamos uma alta cobertura para algumas cidades, como mostrado nas Figuras 2.6a e 2.6d (Nova York). No entanto, como podemos observar nas Figuras 2.6b e 2.6e, o sensoriamento no Cairo, que também possui um número elevado de habitantes, é significativamente mais baixo. Tamanha diferença na cobertura pode ser explicada por diversos fatores. Além dos aspectos econômicos, diferenças na cultura dos habitantes desta cidade, quando comparadas com as culturas presentes nas outras cidades estudadas, podem ter um impacto significativo na adoção e uso desses sistemas considerados [Barth 1969].

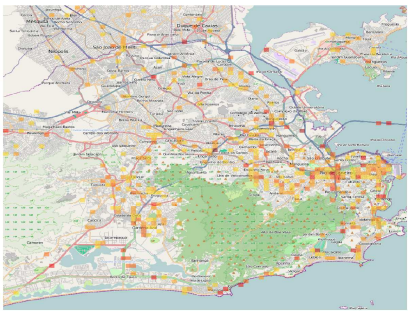


Figura 2.7. Cobertura espacial da RSP para compartilhamento de alerta de trânsito no Rio de Janeiro [Silva et al. 2013e].

Além disso, pode-se observar que a cobertura em algumas cidades, como no Rio de Janeiro (Figuras 2.6c e 2.6f), é bem mais heterogênea quando comparada com a cobertura de Nova York. Isto ocorre, provavelmente, por causa dos aspectos geográficos particulares dessas cidades, ou seja, grandes áreas verdes e grandes porções d'água. O Rio de Janeiro tem a maior floresta urbana do mundo, localizada no meio da cidade, além de muitas colinas de difícil acesso humano. Estes aspectos geográficos limitam a cobertura do sensoriamento. Além disso, os pontos de interesse público, tais como pontos turísticos e centros comerciais, são distribuídos de forma desigual pela cidade. Há grandes áreas residenciais com poucos pontos desse tipo, enquanto outras áreas têm grande concentração dos mesmos.

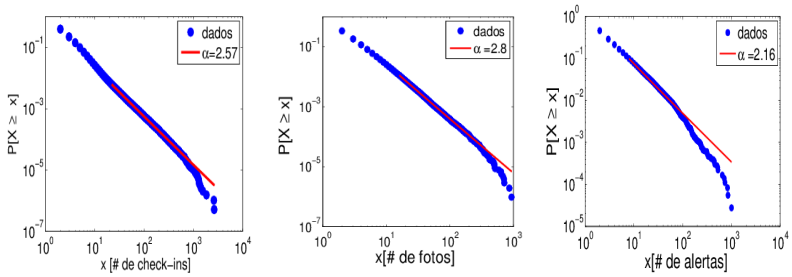
A cobertura espacial dos dados da RSP para alertas de trânsito não é tão abrangente, como das RSPs para compartilhamento de localização e de foto. Isso pode ser observado na Figura 2.7, que mostra o número de alertas em diferentes regiões do Rio de Janeiro por um mapa de calor. Um fator que pode ajudar a explicar isso é a população de usuários do *dataset* de alertas de trânsito, que é menor do que nos outros casos estudados. Outro fator é que os usuários podem ter menos oportunidades para compartilhar alertas de trânsito em comparação com oportunidades para compartilhar fotos ou *check-ins*.

Como a atividade de participação pode ser bastante heterogênea dentro de uma cidade, analisamos a cobertura de RSPs em áreas específicas de uma cidade. Para ter um ID de uma área específica da cidade para os *datasets* do Instagram e Waze, propomos dividir a área das cidades em espaços retangulares menores, como em uma grade[15]. Chamaremos cada área retangular de uma *área específica* dentro de uma cidade. Consideramos que uma área específica possui a seguinte delimitação: $1{\cdot}10^{-4^\circ}$ (latitude) $\times$ $1{\cdot}10^{-4^\circ}$ (longitude). Isso representa uma área de aproximadamente 8×11 metros em Nova Iorque e 10×11 metros no Rio de Janeiro. Para outras cidades, as áreas também podem variar um pouco, mas não a ponto de afetar significativamente as análises realizadas.

A Figura 2.8 apresenta a função de distribuição acumulada complementar (*complementary cumulative distribution function* - CCDF) do número de dados compartilhados (*check-ins*, fotos ou alertas) por área específica de todas as localidades em nossos *datasets*. Primeiramente, observe que, em ambos os casos, uma lei de potência[16] descreve bem

[15]Note que nas áreas selecionadas não são consideradas fronteiras.

[16]Matematicamente, uma quantidade x segue uma lei de potência se ela pode ser obtida de uma distribuição de probabilidade $p(x) \propto x^{-\alpha}$, onde α é um parâmetro constante conhecido como expoente ou parâmetro escalar, e é um valor tipicamente entre $2 < \alpha < 3$ [Clauset et al. 2009].



(a) Foursquare (b) Instagram (c) Waze

Figura 2.8. Distribuição do número de dados em áreas específicas (escala log-log) [Silva et al. 2013a, Silva et al. 2013a, Silva et al. 2013e].

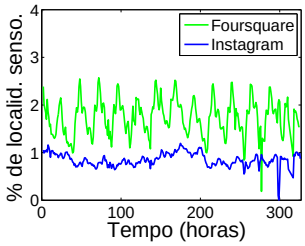


Figura 2.9. Porcentagem de áreas específicas sensoriadas ao longo do tempo [Silva et al. 2014a].

esta distribuição. Isso implica que, na maioria das áreas específicas, há poucos dados compartilhados, enquanto existem algumas poucas áreas com centenas de dados compartilhados. Estes resultados estão consistentes com os resultados apresentados em [Noulas et al. 2011a], trabalho que estudou a participação de usuários em sistemas de compartilhamento de localização. Nos sistemas analisados, é natural que algumas áreas possuam mais atividade que outras. Por exemplo, em áreas turísticas o número de fotos compartilhadas tende a ser maior do que em um supermercado, apesar de um supermercado ser geralmente um local bastante popular. Se uma determinada aplicação requer uma cobertura mais abrangente, é necessário incentivar os usuários a participarem em locais que eles usualmente não o fariam. Micro pagamentos ou sistemas de pontuação são exemplos de alternativas que poderiam funcionar nesse caso. Discutimos essas oportunidades na Seção 2.8.3.

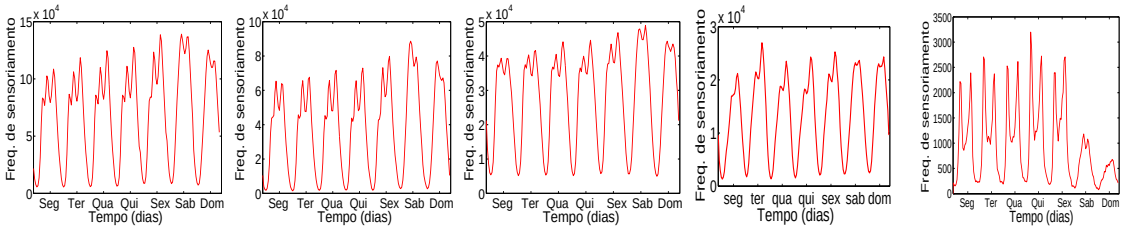
Mostramos que uma RSP pode ter uma cobertura em escala planetária. No entanto, essa cobertura pode ser bastante desigual, em que grandes áreas ficam praticamente descobertas. Com isso em mente, a Figura 2.9 mostra a percentagem de locais distintos onde os usuários compartilharam dados em um determinado intervalo de tempo no Instagram e Foursquare[17], que possuem 598.397 e 725.419 locais, respectivamente. O percentual máximo de locais distintos compartilhados por hora é inferior a 3% para todos os sistemas. Isto indica que a cobertura instantânea destas RSPs é muito limitada quando consideramos todas as localidades que poderiam ser sensoriadas no planeta (considerando todas as localidades já sensoriadas pelo menos uma vez). Em outras palavras, a probabilidade de uma área específica aleatória ser sensoriada em um horário aleatório é bem baixa.

2.5.3. Rotinas e o Compartilhamento de Dados

Analisamos agora como a rotina dos humanos afeta o compartilhamento dos dados. A Figura 2.10 mostra o padrão semanal de compartilhamento de dados em todos os tipos de RSPs analisadas[18]. Como esperado, os dados compartilhados nas RSPs apresentam um padrão diurno, o que implica que durante a madrugada a atividade de sensoriamento é bastante baixa.

[17]Consideramos os *datasets* Instagram2 e Foursquare3, pois representam o mesmo intervalo de tempo.

[18]O horário do compartilhamento foi normalizado de acordo com o local onde o dado foi compartilhado, utilizando para isso a informação geográfica do local.



(a) Foursquare (b) Gowalla (c) Brightkite (d) Instagram (e) Waze

Figura 2.10. Padrão do compartilhamento de fotos durante os dias da semana [Silva et al. 2013a, Silva et al. 2013c, Silva et al. 2013e].

Considerando dias de semana, é possível observar um ligeiro aumento da atividade ao longo da semana, com poucas exceções quando há um pico de atividade. O trabalho [Cheng et al. 2011], que analisou sistemas para compartilhamento de localização, foi observado esse mesmo comportamento, sem nenhum dia como exceção.

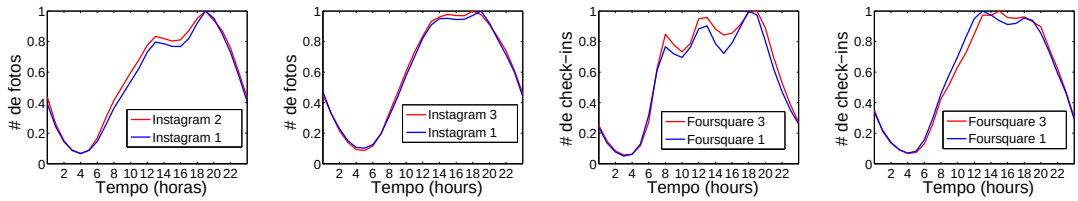
Podemos ainda observar que alguns picos de atividade variam ao longo do dia de acordo com o propósito da RSP. Como podemos ver na Figura 2.10, na RSP para compartilhamento de localizações (Figuras 2.10a–c) existem três picos evidentes por volta da hora do café da manhã, almoço e jantar. Isso também foi observado em [Cheng et al. 2011]. Já na RSP para compartilhamento de fotos (Figura 2.10d) existem apenas dois picos evidentes, que ocorrem por volta da hora do almoço e jantar. E no caso da RSP para compartilhamento de alertas de trânsito (Figura 2.10e) também existem dois picos evidentes, um por volta de 7:00 e 8:00 da manhã e outro por volta de 6:00 da tarde, coincidindo com horários típicos de maior intensidade no trânsito.

Analisando os diferentes padrões de comportamento para dias de semana e final de semana podemos observar que o padrão é significativamente diferente. Note que os picos observados nos dias de semana não são evidentes nos finais de semana. A falta de rotina bem definida nos fins de semana é uma das possíveis explicações para esse fato. Além disso, as diferenças entre dias de semana e final de semana possuem relação com o tipo de sistema analisado. Por exemplo, como nos fins de semana muitas pessoas não precisam dirigir, é natural esperar um volume menor de dados no Waze.

A Figura 2.11 mostra o padrão temporal de compartilhamento para o Instagram e o Foursquare considerando todos os *datasets*. Essa figura apresenta o número médio de dados compartilhados por hora durante, durante os dias de semana (de segunda a sexta-feira) e também durante o fim de semana (sábado e domingo). Surpreendentemente, vemos o mesmo padrão de compartilhamento para cada curva é muito semelhante, apesar do enorme intervalo entre as coletas (aproximadamente um ano). Isso acontece para os dias de semana e fins de semana, sugerindo que o comportamento do usuário em ambos os sistemas tende a se manter consistente ao longo do tempo. Esse é um resultado interessante e importante, pois mostra que podemos usar diferentes *datasets* para propósitos similares.

Mostramos agora como as rotinas impactam no comportamento de compartilhamento durante a semana. Para essa análise, consideramos os *datasets* do Instagram e Foursquare para Nova York, São Paulo e Tóquio. Os resultados são mostrados na Figura 2.12[19].

[19]Cada curva é normalizada pelo número máximo de conteúdo compartilhado em uma região específica

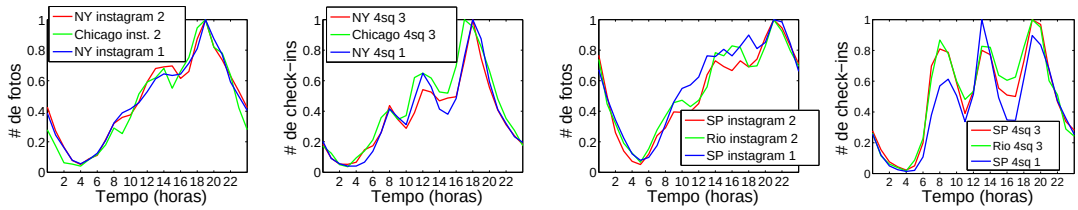


(a) Instagram – dia de semana (b) Instagram – fim de semana (c) Foursquare – dia de semana (d) Foursquare – fim de semana

Figura 2.11. Padrão de compartilhamento temporal no Instagram e Foursquare [Silva et al. 2013d].

Em todas as figuras nós exibimos dados dos *datasets* do mesmo período (Instagram2 e Foursquare3) para duas cidades do mesmo país, e dados de um *dataset* com período anterior (Instagram1 e Foursquare1) para uma dessas cidades, como uma referência de comparação.

Primeiramente, observe a distinção entre as curvas de cada cidade no mesmo sistema (por exemplo, Instagram, Figuras 2.12a, c, e) e também em diferentes sistemas (por exemplo, as Figuras 2.12a e 2.12b para Nova Iorque). Em seguida, observe que o padrão de compartilhamento para cada cidade no mesmo país é bastante semelhante, o que pode ser consequência dos padrões culturais dos habitantes desses países. Isso representa, de certa maneira, uma assinatura de aspectos culturais, o que ilustra, mais uma vez, o potencial desse tipo de dado para o estudo de dinâmica de cidades e do comportamento social urbano.



(a) Nova Iorque – Instagram (b) Nova Iorque – Foursquare (c) São Paulo – Instagram (d) São Paulo – Foursquare

Figura 2.12. Padrão de compartilhamento temporal do Instagram e Foursquare para Nova Iorque, São Paulo e Tóquio durante dias de semana [Silva et al. 2013d].

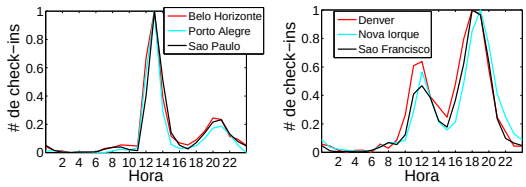
Podemos ainda analisar classes de locais específicos. As Figuras 2.13a e 2.13b[20] mostram o número de *check-ins* realizados em restaurantes ao longo das horas do dia, durante os dias de semana, em diferentes cidades do Brasil e dos Estados Unidos. Estes resultados capturaram diferenças importantes entre as culturas dos dois países: enquanto o jantar é a refeição principal para os americanos, o almoço desempenha um papel mais importante nos hábitos alimentares dos brasileiros.

2.5.4. Comportamento dos Nós

Nesta seção analisamos o desempenho dos nós da RSP (i.e., dos usuários) quanto ao compartilhamento de dados. A Figura 2.14 mostra a distribuição do número de dados (fotos e

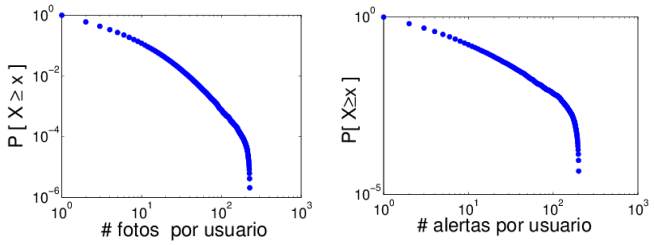
representando a cidade.

[20]Os valores são normalizados pelo valor máximo encontrado em qualquer hora para a cidade específica.



(a) Cidades brasileiras (b) Cidades dos EUA

Figura 2.13. Número médio de *check-ins* em restaurantes durante dias de semana ao longo das horas do dia [Silva et al. 2014a].



(a) Instagram (b) Waze

Figura 2.14. Distribuição do número de dados compartilhadas pelos usuários [Silva et al. 2013c, Silva et al. 2013e].

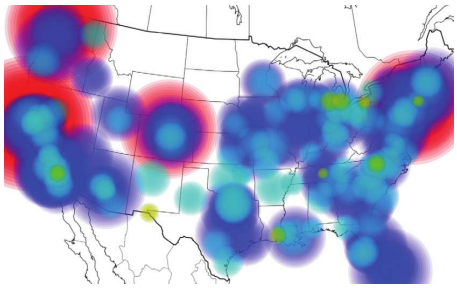


Figura 2.15. Resultado da métrica raio de giro [Cheng et al. 2011].

alertas) compartilhados por cada usuário da nossa base de dados. Como podemos observar, a distribuição possui cauda pesada, o que significa que a participação dos usuários pode ser muito desigual. Por exemplo, aproximadamente 40% dos usuários contribuíram com apenas uma foto no período considerado, enquanto somente 17% e 0,1% dos usuários contribuíram com mais que 10 e 100 fotos, respectivamente. É natural que essa variabilidade aconteça por diversos motivos. Por exemplo, alguns usuários podem dar mais importância para quesitos de privacidade do que outros. Uma cauda pesada também é observada na distribuição do número de *check-ins*, como foi mostrado em [Noulas et al. 2011a]. Cerca de 20% dos usuários realizaram apenas um *check-in*, 40 % acima de 10, ao passo que cerca de 10 % realizaram mais de 100 *check-ins*.

Além disso, em [Silva et al. 2013c, Silva et al. 2013e, Silva et al. 2013a] mostramos que há momentos em que muitos dados são compartilhados em intervalos de poucos minutos e momentos em que não há compartilhamento por horas. Isso pode indicar que a maioria do compartilhamento de dados acontece em intervalos específicos, provavelmente relacionados ao ciclo circadiano (ou rotina) das pessoas. Por exemplo, o compartilhamento de fotos em restaurantes tende a acontecer mais nos horários de almoço e jantar. Aplicações baseadas nesse tipo de sensoriamento devem considerar que a participação do usuário pode variar significativamente ao longo do tempo.

Observamos ainda que uma fatia significativa dos usuários realiza compartilhamento consecutivo de fotos num curto intervalo de tempo. Por exemplo, cerca de 20% de todo o compartilhamento de fotos observado acontece em até 10 minutos. Isso sugere que os usuários tendem a compartilhar mais de uma foto na mesma área. Em [Noulas et al. 2011a] os autores também observaram que uma parcela significativa dos *check-ins* no Foursquare

são realizados dentro de um curto intervalo de tempo. Por exemplo, mais do que 10% de *check-ins* ocorrem dentro de 10 minutos.

Em [Cheng et al. 2011] os autores analisaram *check-ins* compartilhados em vários serviços de compartilhamento de localização. Eles descobriram que os usuários possuem padrões simples e reproduzíveis, e também que o status social, além de fatores geográficos e econômicos, colaboram com a mobilidade.

Para fazer essa análise os autores usaram três propriedades estatísticas para estudar e modelar padrões de mobilidade humana: *deslocamento* (*displacement*); *raio de giro* (*radius of gyration*); e *probabilidade de retorno* (*returning probability*). Para ilustrar um de seus resultados, a Figura 2.15 mostra o raio médio de giro dos usuários em grandes cidades (com mais de 100.000 habitantes) nos EUA. As bolhas vermelhas[21] são cidades com um raio de giro maior do que 500 milhas; as azuis são cidades com um raio maior do que 250 milhas; as de cor ciano possuem um raio maior do que 125 milhas; e as amarelos são o resto das grandes cidades analisadas. Usuários em cidades costeiras tendem a ter um raio maior de giro do que os usuários em cidades do interior, e as pessoas em estados centrais tendem a ter um alto raio de giro devido a viagens de longa distância para o litoral [Cheng et al. 2011].

Da mesma forma, em [Cho et al. 2011] os autores investigaram padrões de movimentos e como os laços sociais podem impactar nesses movimentos. Os autores observaram que viagens de curta distância são espacialmente e temporalmente periódicas e não são afetadas pela estrutura de rede social, enquanto as viagens de longa distância são mais influenciadas por laços da rede social.

2.5.5. Considerações Finais

Identificamos várias propriedades de RSPs em comum: (i) escala planetária; (ii) frequência altamente desigual de compartilhamento de dados, tanto espacialmente quanto temporalmente, o que é altamente correlacionado com a rotina típica das pessoas; (iii) a participação do usuário em relação ao número de dados compartilhados e onde esses dados são compartilhados pode variar significativamente; (iv) o padrão temporal de compartilhamento não varia consideravelmente ao longo do tempo para o mesmo tipo de sistema.

As propriedades identificadas revelam o potencial de RSPs para conduzir vários estudos sobre a dinâmica da cidade e do comportamento social urbano. Além disso, o entendimento do comportamento do usuário é o primeiro passo para modelá-lo. Com modelos que explicam o comportamento do usuário podemos fazer previsões de ações e desenvolver melhores sistemas para planejamento de capacidade de carga do sistema.

Na Seção 2.7.10 são discutidas algumas das principais tecnologias e ferramentas para a análise de dados de RSPs, que podem ser bastante úteis em futuras análises de outros dados.

2.6. Aplicações e Serviços Relacionados ao Estudo de Sociedades Urbanas

Nesta seção, discutiremos as abordagens e modelos utilizados em diversas aplicações e serviços relacionados ao estudo da dinâmica da cidade e do comportamento social urbano utilizando dados de RSPs. Para a construção de novos serviços e aplicações nessa área é de

[21] Consultar a versão digital disponível online em cores.

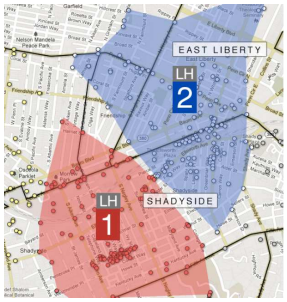


Figura 2.16. "Livehoods" encontrados em Nova Iorque [Cranshaw et al. 2012].

suma importância conhecer as propriedades dos dados da RSP em estudo.

Os estudos mostrados aqui foram agrupados em cinco classes: Funcionamento de Cidades (Seção 2.6.1); Mobilidade Urbana (Seção 2.6.2); Padrões Sociais, Econômicos e Culturais (Seção 2.6.3); Detecção de Eventos e Interesses (Seção 2.6.4); e Problemas das Cidades (Seção 2.6.5).

2.6.1. Funcionamento de Cidades

As informações obtidas a partir de RSPs têm o poder de mudar os nossos limites físicos percebidos, bem como ajudar a compreender melhor a dinâmica de cidades. Esta seção concentra na apresentação de estudos nessas direções.

Usando dados do Foursquare, em [Cranshaw et al. 2012] os autores propuseram um modelo para identificar regiões distintas de uma cidade que refletem padrões atuais de atividades coletivas, apresentando novos limites para os bairros. A ideia é expor a natureza dinâmica das áreas urbanas locais, considerando a proximidade espacial (derivado de coordenadas geográficas) e proximidade social (derivado da distribuição de check-ins) de locais.

Para isso, os autores utilizaram dados do Foursquare e desenvolveram um modelo que agrupa locais semelhantes considerando características sociais e espaciais. Cada *cluster* representa diferentes fronteiras geográficas dos bairros. O método de agrupamento utilizado é uma variação do agrupamento espectral proposto por [Ng et al. 2002].

A Figura 2.16 mostra dois *clusters* (ou "*livehoods*", nome usado pelos autores), encontrados em Nova Iorque, representados pelos números 1 e 2. Nessa figura as linhas pretas indicam os limites oficiais da cidade. Veja que os limites dos *clusters* são bastante diferentes. Para tentar validar esses resultados os autores usaram resultados de entrevistas com moradores da cidade. De acordo com as respostas coletadas, esses e outros *clusters* eram esperados.

Em [Noulas et al. 2011b] os autores propuseram uma abordagem para classificar áreas e usuários de uma cidade usando categorias de locais do Foursquare. Isso poderia ser usado para identificar as comunidades de usuários que visitam categorias semelhantes de lugares, útil para sistemas de recomendação, ou na comparação de áreas urbanas dentro e entre as cidades. A abordagem utilizada é baseada em algoritmo de agrupamento espectral [Ng et al. 2002].

Mais especificamente, os autores consideram a atividade dos usuários do Foursquare

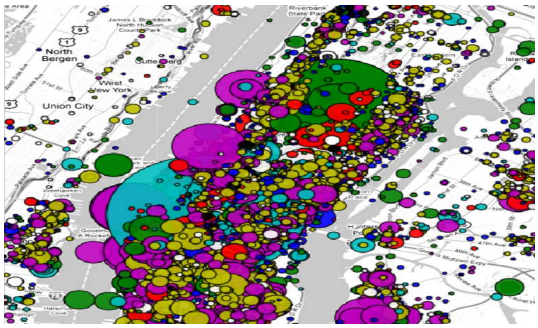


Figura 2.17. Atividade dos usuários do Foursquare para Nova Iorque. A&E (vermelho); Edu (preto); Outd (verde); NL (magenta); Shop (branco); and Trvl (ciano) [Noulas et al. 2011b].

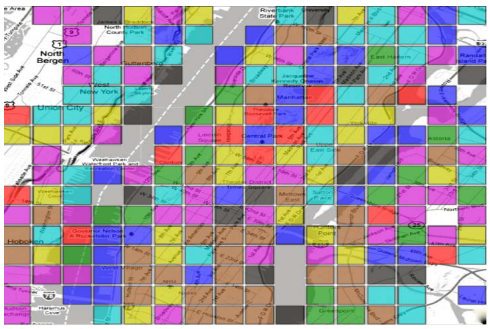


Figura 2.18. Visualização do agrupamento espectral. Cada cor simboliza um *cluster*. [Noulas et al. 2011b].

para Nova Iorque, como mostrado pela Figura 2.17. Nessa figura, um círculo representa um local e seu raio a popularidade em termos de número de checkins. Cada cor corresponde a uma das oito categorias gerais introduzidas pelo Foursquare. Essa figura destaca a diversidade da atividade humana sobre a área considerada.

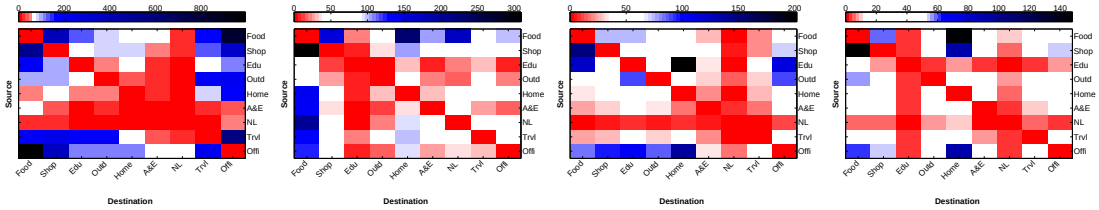
Em seguida, os autores dividiram uma cidade para ser analisada em áreas de tamanhos iguais, cada uma delas será um dado de entrada para o algoritmo de agrupamento. Para cada área é calculada a atividade realizada pelos usuários com base nas visitas em locais dessa região. Com isso, calcula-se a semelhança entre duas áreas utilizando a similaridade do cosseno (*cosine similarity*), entre as atividades representadas. Após esse processo, os autores realizam um agrupamento espectral. O resultado desse processo para a cidade de Nova Iorque é mostrado na Figura 2.18.

Em [Silva et al. 2014d] propusemos uma técnica chamada *City Image*, que fornece um resumo visual da dinâmica da cidade com base nos movimentos das pessoas. Esta técnica explora grafos de transição urbana para mapear os movimentos dos usuários entre locais da cidade. O grafo de transição urbana considerado é um grafo dirigido ponderado $G(V,E)$, em que um nó $v_i \in V$ é a categoria de um local específico (por exemplo, $food$) e uma aresta direcionada $(i,j) \in E$ marca uma transição entre duas categorias. Ou seja, uma aresta existe a partir do nó v_i para o nó v_j se pelo menos um usuário compartilhou um dado em um local categorizado por v_j logo após compartilhar um dado em um local categorizado por v_i. O peso $w(i,j)$ de uma aresta é o número total de transições que ocorreram a partir de v_i para v_j. Somente dados consecutivos compartilhados pelo mesmo usuário dentro de 24 horas, com início às 5:00, são considerados no cálculo de uma transição.

A *City Image* é uma técnica promissora que permite uma melhor compreensão da dinâmica de cidades, ajudando na visualização das rotinas comuns de seus cidadãos. Cada célula da *City Image* representa o quão favorável é uma transição de uma determinada categoria em um determinado lugar (eixo vertical) para outra categoria (eixo horizontal), valores que são calculados utilizando um modelo aleatório/nulo [Silva et al. 2014d]. As cores vermelhas representam rejeição, as cores azuis representam favorabilidade e a cor branca representa indiferença. Nós exemplificamos a técnica *City Image* para duas cidades[22]: São Paulo

[22]Utilizando dados do *dataset* Foursquare1.

(Figures 2.19a and 2.19b); e Kuwait (Figures 2.19c e 2.19d). Para ambos os casos, consideramos dias de semana durante o dia, que é o período típico de rotinas, e fim de semana durante a noite, que é um período representativo de atividades de lazer (fora da rotina).



(a) SP (Dia - Dia de semana) (b) SP (Noite - Fim de semana) (c) KU (Dia - Dia de semana) (d) KU (Noite - Fim de semana)

Figura 2.19. Images produzidas com a técnica *City Image* para São Paulo (SP) e Kuwait (KU) em diferentes períodos. Abreviaturas das categorias de locais (Nomes usados pelo Foursquare): *Arts & Entertainment* (A&E); *College & Education* (Edu); *Great Outdoors* (Outd); *Nightlife Spot* (NL); *Shop & Service* (Shop); and *Travel Spot* (Trvl) [Silva et al. 2014a].

Primeiramente, observe que transições para *office* (locais de trabalho) são mais prováveis de acontecer nos dias de semana e durante o dia para ambas as cidades, como esperado. No entanto, note que as imagens da cidade de São Paulo e Kuwait também têm diferenças significativas que refletem diversidades culturais entre ambas as cidades. Note, por exemplo, que a imagem que representa transições nas noites do fim de semana (Figure 2.19d) mostra a falta de transições favoráveis para a categoria *nightlife* no Kuwait. Este não é o caso de São Paulo (Figure 2.19b), em que a transição *food* $\rightarrow$ *nightlife* é altamente favorável de acontecer. Isso sugere que em São Paulo as pessoas gostam de frequentar locais relacionados com comida (*food*) antes de ir para casas noturnas (*nightlife*). No Kuwait, em vez disso, as pessoas são provavelmente mais favoráveis a realizarem as transições *shop* $\rightarrow$ *food* e *food* $\rightarrow$ *home* nas noites do fim de semana.

Técnicas para facilitar a interpretação das rotinas de habitantes de uma cidade, tais como as mencionadas aqui, são ferramentas valiosas para ajudar os urbanistas a entender melhor a dinâmica de cidades e, consequentemente, tomar decisões mais eficazes em relação à problemas das cidades, por exemplo.

2.6.2. Mobilidade Urbana

Nesta seção apresentamos trabalhos que se concentram em estudar padrões de mobilidade urbana dos usuários com dados de RSPs. Esses dados incluem informações espaço-temporais, por exemplo, *check-ins* e fotos com coordenadas geográficas. O estudo da mobilidade é útil para muitas finalidades. Com dados de RSPs é possível entender, por exemplo, como os usuários alocam tempo para diferentes atividades, sendo, portanto, uma questão fundamental nas ciências sociais. Além disso, é possível projetar novas ferramentas para ajudar os engenheiros de tráfego a entender o fluxo de pessoas na cidade.

A modelagem dos padrões de mobilidade vem atraindo a atenção de pesquisadores em diferentes áreas, como física e computação [Brockmann et al. 2006, Zheng et al. 2009, Gonzalez et al. 2008]. É importante ressaltar que os dados derivados das RSPs são diferentes de dados provenientes de *traces* de GPS ou de dados tradicionais do uso do telefone

(a) Caminho mais curto (b) Caminho mais bonito

Figura 2.20. Mapas mostrando diferentes caminhos entre os mesmos locais [Quercia et al. 2014].

celular, como ligações telefônicas, e apresentam características especiais e variados contextos. Por exemplo, os *check-ins* em serviços de compartilhamento de localização ou fotos em um serviço de compartilhamento de fotos trazem informações extras sobre um lugar particular. Por exemplo, um *check-in* está associado com um tipo de local, e.g. bar, e uma foto pode trazer informações sobre a situação atual dentro deste local. Com isso, nosso foco aqui são estudos que analisam dados de RSPs.

Em [Quercia et al. 2014] os autores propuseram uma metodologia para recomendação de rotas que leva em consideração não somente o menor caminho, mas também características emocionais, por exemplo, beleza. Nem sempre o menor caminho é o que gostaríamos de percorrer. Um turista, por exemplo, pode optar por um caminho mais bonito, mesmo que a distância seja um pouco maior. A Figura 2.20 mostra dois caminhos entre os mesmos locais na cidade de Londres, em que um é o mais curto (Figura 2.20a) e o outro o mais bonito (Figura 2.20b).

Para quantificar o quão localidades urbanas são agradáveis, os autores usaram dados de uma plataforma de *crowd-sourcing* que mostra duas cenas de ruas em Londres (considerando centenas), e um usuário vota em qual acha mais bonita, tranquila e feliz. Em seguida, os autores traduzem os votos em medidas quantitativas de percepção de localização. Depois disso, os autores criam um grafo considerando essas localizações. Para isso, os autores dividiram a área da cidade em células de 200x200 metros. Cada célula é um nó no grafo e cada nó possui arestas com nós que representam células vizinhas. Esse grafo permite a descoberta de caminhos agradáveis.

Nguyen and Szymanski [Nguyen and Szymanski 2012] usaram dados do Gowalla para criar e validar modelos de mobilidade e relações humanas. Nesse trabalho, os autores propuseram um modelo de mobilidade baseado em amizade (FMM), que leva em conta os laços sociais, a fim de fornecer um modelo mais preciso da mobilidade humana. Com esse modelo, os autores foram capazes de estudar a frequência com que amigos viajam juntos. Ele pode melhorar a precisão de um número variado de aplicações, tais como engenharia de tráfego em redes de comunicação, sistemas de transporte e planejamento urbano.

O modelo de mobilidade proposto utiliza um modelo de Markov, onde os estados representam locais de *check-ins* e as ligações representam a probabilidade de ir de um lugar para outro. Por exemplo, a probabilidade de ir do trabalho para bar é definida como a razão entre o número de vezes que um determinado usuário executa um *check-in* em um bar logo após realizar um *check-in* no trabalho, e o número de vezes que o usuário realiza um *check-in* no trabalho.

Em [Zheng et al. 2012] os autores estudaram a mobilidade e padrões de viagem de turistas a partir de fotos compartilhadas no Flickr. A fim de extrair os padrões de viagem, os autores focaram as análises no movimento de turistas de acordo com as regiões atrativas e características topológicas de rotas de viagem feitas por diferentes turistas. Para isso, primeiro é construído um banco de dados de caminhos turísticos com base no conceito de entropia de mobilidade (considerando entropia de Shannon [Shannon 1948]), usada para discriminar o movimento turístico do não turístico.

Em seguida, os autores propõem um método para descobrir regiões atrativas em uma cidade, usando para isso o algoritmo de agrupamento DBSCAN [Ester et al. 1996]. Para estudar o movimento turístico, os autores consideraram um modelo de Markov criado a partir da sequência de visitas nas regiões atrativas. Com isso, os autores podem estimar as estatísticas de visitantes que viajam de uma região para outra. Para estudar as características topológicas de rotas de turismo, os autores realizam um agrupamento de rotas de viagem, aplicando uma versão modificada da maior subsequência comum (*longest common subsequence*), como uma métrica de similaridade para minimizar o ruído.

Esses esforços ilustram o crescente interesse e o potencial de utilização de dados compartilhados em RSPs para estudar padrões de mobilidade de humanos em larga escala.

2.6.3. Padrões Sociais, Econômicos e Culturais

Os dados de RSPs também podem ser usados para estudar aspectos sociais, econômicos e culturais dos habitantes de cidades. Por exemplo, pode-se argumentar que uma pequena quantidade de dados compartilhados em uma área da cidade pode indicar uma falta de acesso à tecnologia por parte da população local, pois o uso de serviços de compartilhamento de localização muitas vezes dependem de *smartphones* e planos de dados 3G ou 4G que, geralmente, são caros. Nessa direção, em [Silva et al. 2013a] nós mostramos que a análise de dados de RSPs permitem a visualização de fatos interessantes relacionados com questões socioeconômicas de uma cidade. Por exemplo, dados de uma RSP para compartilhamento de localização para a cidade do Rio de Janeiro são escassos em áreas pobres, incluindo as que são localizadas muito perto de áreas ricas. Essa informação pode ser útil para gerar melhores políticas públicas nessas áreas. Note que a mesma informação pode ser obtida utilizando métodos tradicionais, tais como questionários, mas esse processo é muito mais lento e caro.

Com o intuito de melhor entender padrões sociais a partir da análise de dados de RSPs, em [Quercia et al. 2012] os autores estudaram como comunidades virtuais, observadas nos sistemas analisados, se assemelham às comunidades da vida real. Os autores testaram se teorias sociológicas estabelecidas de redes sociais da vida real são válidas nessas comunidades virtuais. Eles descobriram, por exemplo, que os influentes (*social brokers*) no Twitter são líderes de opinião que se arriscam compartilhando *tweets* sobre diferentes temas. Eles também descobriram que a maioria dos usuários têm redes geograficamente locais, e que os influentes expressam não apenas emoções positivas, mas também negativas.

Para realizar este trabalho, os autores aplicaram métricas de rede que a literatura afirma que podem estar relacionadas com relações sociais, como a reciprocidade e restrição da rede [Quercia et al. 2012]. A reciprocidade r é a proporção de arestas em uma rede que são bidirecionais $L^{<->}$ em relação ao número total de arestas L: $r = \frac{L^{<->}}{L}$. Considerando

uma rede social focada em um vértice (“ego”) e vértices e arestas a quem o ego está diretamente conectado, valores baixos de reciprocidade poderiam indicar, por exemplo, uma rede social de uma celebridade. A restrição da rede mede as oportunidades de se tornar influente (*brokerage opportunities*). Um alto valor de restrição da rede significa menos oportunidades. Os autores usaram a formulação de Burt [Burt 1992] nesse caso específico. Mais detalhes sobre essas e outras métricas de rede que podem ser usadas para análises de relações sociais podem ser encontradas no livro: [Easley and Kleinberg 2010].

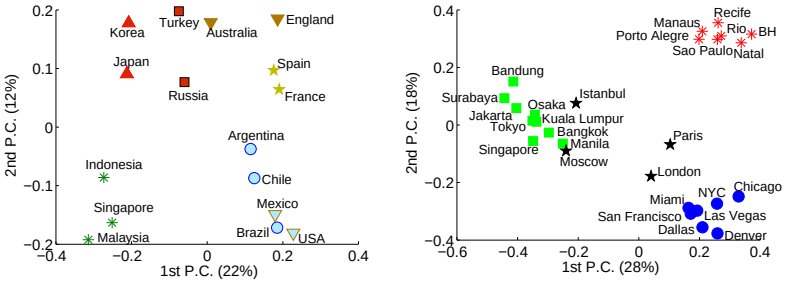
Numa direção similar, em [Joseph et al. 2012] os autores analisaram um conjunto de dados do Foursquare para identificar grupos de pessoas e os lugares que elas visitam. O modelo utilizado foi capaz de identificar grupos de pessoas que representam grupos espacialmente próximos e pessoas que parecem ter interesses semelhantes.

A abordagem utilizada se baseia na ideia de modelos probabilísticos para tópicos. Para isso, eles usaram o modelo de alocação latente de Dirichlet (LDA - *Latent Dirichlet Allocation*) [Blei et al. 2003], que é, geralmente, utilizado para estudar documentos. Na instanciação do modelo, cada *check-in* de um usuário é encarado como uma palavra de um “documento” que representa um usuário, isso de forma análoga a documentos de texto, onde um documento pode ter muitas palavras.

Além disso, ao estudar o comportamento social de áreas específicas, uma das primeiras perguntas que surgem é: o quão diferente uma cultura é de outra? Sabemos que os hábitos alimentares e de bebidas são capazes de descrever fortes diferenças culturais. Com base nisso, em [Silva et al. 2014c] propomos uma nova metodologia para a identificação de fronteiras culturais e semelhanças entre sociedades, considerando hábitos alimentares e de bebida. Para isso, foram usados *check-ins* do Foursquare para representar as preferências do usuário em relação ao que se come e bebe localmente, por exemplo, em uma determinada cidade.

Essa análise surpreendentemente diz muito sobre as diferenças e semelhanças entre as culturas. Para isso estudamos a correlação entre os *check-ins* dados em diferentes tipos de restaurantes para várias cidades ao redor do mundo. Observamos que as cidades de um mesmo país, onde os habitantes normalmente possuem cultura e hábitos alimentares semelhantes, têm as correlações mais fortes com relação às preferências de restaurante. Além de preferências para as categorias de alimentos, também podemos ver diferenças nos horários em que as pessoas vão a restaurantes e compartilham dados, como foi apresentado na Seção 2.5.3. Essas análises permitiram a proposição de uma metodologia para a identificação de culturas semelhantes, que pode ser aplicada em regiões de tamanhos variados, como países, cidades ou até mesmo bairros [Silva et al. 2014c]. Nessa metodologia é utilizado um algoritmo de agrupamento baseado em particionamento ($k-means$ [Hartigan and Wong 1979]), bem como a técnica de análise de componentes principais [Jolliffe 2002]. Os resultados para países e cidades são ilustrados nas Figuras 2.21a e 2.21b, mostrando como culturas semelhantes são bem separadas. Nessas figuras foram usados os dois principais componentes apenas para mostrar os resultados, no entanto a obtenção do resultado considerou todos os componentes.

As diferenças culturais utilizando dados de RSPs também foram estudadas em [Hochman and Schwartz 2012], que investigaram as preferências de cores em fotos compartilhadas no Instagram. Os autores encontraram diferenças consideráveis entre imagens



(a) Países (b) Cidades

Figura 2.21. Grupos encontrados utilizando a metodologia de separação de culturas. Cada símbolo reflete um grupo [Silva et al. 2014c].

Figura 2.22. Localização das lojas analisadas [Karamshuk et al. 2013].

de países com culturas distintas. Na mesma direção, em [Poblete et al. 2011] os autores investigaram como o comportamento de divulgação de conteúdo no Twitter varia entre alguns países, bem como as possíveis explicações para essas diferenças. A investigação das distinções culturais entre diferentes cidades e países é valiosa em muitas áreas e pode auxiliar várias aplicações. Por exemplo, como cultura é um aspecto importante por razões econômicas, a identificação de semelhanças entre os lugares que estão geograficamente separados pode ser necessária para empresas que possuem negócios em um país e querem avaliar a compatibilidade de preferências entre diferentes mercados.

Relacionado com o aspecto econômico das cidades, em [Karamshuk et al. 2013] os autores estudaram o problema da alocação ótima de lojas de varejo na cidade. Eles usaram dados do Foursquare para compreender como a popularidade de três redes de lojas de varejo em Nova Iorque (veja a Figura 2.22) é definida em termos de número de *check-ins*.

Foram avaliadas um conjunto diversificado de características (*features*), modelando informações espaciais e semânticas sobre os locais e padrões de movimentos dos usuários na área ao redor do local analisado. Os autores observaram que a presença de locais que atraem muitos usuários naturalmente, como estação de trem ou aeroporto, bem como lojas de varejo do mesmo tipo das analisadas, que definem a concorrência comercial local de um área, são os indicadores mais fortes de popularidade.

2.6.4. Detecção de Eventos e Interesses

A identificação de eventos e pontos de interesse através de dados de RSPs é beneficiada pela natureza de tempo (quase) real das RSPs. Eventos podem ser naturais, tais como terremotos, ou não naturais, tais como a identificação/previsão de mudanças no mercado de ações. Por sua vez, um ponto de interesse é uma localização específica que alguém pode achar útil ou interessante, como um restaurante ou um estádio de futebol.

Em relação à detecção de eventos, em [Gomide et al. 2011] os autores analisaram como a epidemia de Dengue é refletida no Twitter e em que medida essa informação pode ser usada na vigilância dessa doença. Os autores mostraram que o Twitter pode ser usado para prever, espacial e temporalmente, epidemias de dengue. Eles analisam como dados do Twitter refletem epidemias com base em quatro dimensões: volume, localização, tempo e percepção do público. Especificamente, os autores estudam como os usuários se referem à

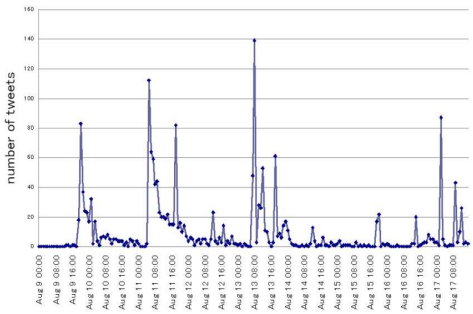


Figura 2.23. Número de *tweets* relacionados com terremotos [Sakaki et al. 2010b].

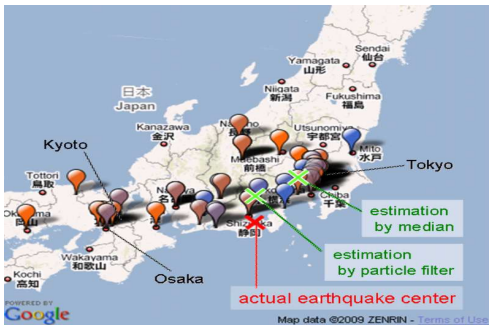


Figura 2.24. Estimativa da localização do terremoto [Sakaki et al. 2010b].

dengue no Twitter com análise de sentimentos [Gonçalves et al. 2013] e usam o resultado para focar em apenas *tweets* que, de alguma forma, expressam experiência pessoal sobre a dengue. Em seguida, os autores construíram um modelo baseado em regressão linear [Yan 2009] para a previsão do número de casos de dengue. O resultado dessa pesquisa é aplicado no projeto chamado Observatório da Dengue[23].

Em [Sakaki et al. 2010b] os autores estudaram a interação em tempo real de acontecimentos no Twitter, por exemplo, terremotos, e propuseram um algoritmo para monitorar mensagens no Twitter para detectar a ocorrência de eventos. Para demonstrar a eficácia de seu método, os autores construíram um sistema de aviso de terremoto no Japão, que foi capaz de detectar 96% dos terremotos relatados pela Agência Meteorológica do Japão (JMA) com escala de intensidade sísmica de 3 ou mais. A notificação foi capaz de ser entregue mais rápida do que os avisos que são transmitidos pela JMA. A abordagem utilizada é um classificador de *tweets* com base em características, tais como: palavras-chave em um *tweet*; o número de palavras; e o seu contexto. Depois disso, os autores produziram um modelo espaço-temporal probabilístico para o evento alvo que pode encontrar o centro e a trajetória do local do evento.

Em os autores [Bollen et al. 2011] estudaram se os estados coletivos de humor derivados de mensagens do Twitter são correlacionados com o valor da bolsa Dow Jones ao longo do tempo. Seus resultados indicam que é possível obter uma boa precisão na previsão das mudanças diárias de alta e queda dos valores de fechamento dessa bolsa de valores. Isso é possível escolhendo dimensões de humor específicos, mas não todos os que foram considerados.

Além de eventos que tendem a acontecer esporadicamente, toda cidade possui um conjunto de áreas que desperta um maior interesse dos residentes ou visitantes, as aqui denominadas *pontos de interesse* (PDI). Dentre os PDIs mais visitados, podemos mencionar os pontos turísticos da cidade. No entanto, nem todos os PDIs de uma cidade são pontos turísticos. Por exemplo, uma área de bares pode ser bastante popular entre os residentes da cidade, mas sem atrativos para turistas. Além disso, PDIs são dinâmicos, ou seja, áreas que são populares hoje podem não ser mais amanhã. Assim, uma aplicação que emerge naturalmente a partir da análise de dados de algumas RSPs, por exemplo para compartilhamento

[23]http://www.observatorio.inweb.org.br/dengue.

de fotos ou localização, é a identificação de PDIs. Isso é possível porque cada foto ou *check-in* representa, implicitamente, um interesse de um indivíduo em um determinado instante. Com isso, quando muitas fotos de um determinado local são compartilhadas dentro de um certo intervalo de tempo, esse local pode ser um PDI.

Uma vantagem de usar RSPs para identificar pontos de interesse na cidade é que podemos obter resultados robustos a mudanças dinâmicas. Ou seja, pelo fato das RSPs fornecerem dados dinâmicos, elas podem capturar automaticamente as alterações nos interesses das pessoas ao longo do tempo, ajudando a identificar rapidamente as áreas que por ventura se tornem um PDI (por exemplo, devido à abertura de um novo restaurante) ou que deixem de ser populares.

A identificação de pontos de interesse em uma cidade foi investigada em [Crandall et al. 2009], onde os autores mostraram como inferir a localização de uma foto sem usar os dados geoespaciais. Na mesma direção, em [Kisilevich et al. 2010] os autores usaram fotos geolocalizadas para analisar e comparar eventos temporais que aconteceram em uma cidade, e também para classificar locais turísticos.

Além disso, em [Silva et al. 2013b] nós também apresentamos uma técnica para identificar PDIs e, a partir deles, identificar pontos turísticos. A técnica considera que cada par i de coordenadas (longitude, latitude) $(x,y)_i$ está associada a um ponto p_i, que representa um dado compartilhado, e.g. uma foto. Nós começamos calculando distância geográfica entre cada par de pontos (p_i, p_j) (usando a fórmula de Haversine [Sinnott 1984]) e agrupamos todos os pontos p_i que estão próximos uns dos outros (utilizando um método de agrupamento hierárquico aglomerativo, usando como critério de ligação: "complete-linkage" [Sørensen 1948, Kaufman and Rousseeuw 2009]). Para capturar os PDIs, usamos um modelo nulo para excluir grupos que possam ter sido gerados por situações aleatórias (ou seja, movimentos de pessoas aleatórias), e, portanto, não refletem a dinâmica da cidade. Para identificar os grupos, analisamos o número de compartilhamento de dados em cada um deles e usamos métodos estatísticos simples. Em seguida, separamos os pontos turísticos dos PDIs assumindo que turistas possuem rotas conhecidas na cidade (mais detalhes em [Silva et al. 2013b]).

Quando aplicada para a cidade de Belo Horizonte considerando dados do Foursquare e Instagram, essa técnica foi capaz de encontrar a maioria dos seus PDIs e pontos turísticos. Os resultados também mostram que diferentes RSPs podem fornecer dados complementares, pois nenhuma RSP encontrou todos os pontos turísticos. Tais diferenças podem refletir mudanças na cidade durante o intervalo de tempo em que um *dataset* específico foi coletado. Por exemplo, durante a coleta do *dataset* Instagram1, Belo Horizonte não estava recebendo jogos de futebol. Isso explica por que o estádio de futebol não foi identificado como um PDI utilizando esse *dataset*. Por outro lado, a análise de um *dataset* do mesmo sistema coletado mais recentemente (Instagram2), identificou corretamente o estádio como um ponto turístico importante da cidade. Isso ilustra como os dados de RSPs podem capturar automaticamente alterações da dinâmica da cidade, sendo úteis para detectar locais incomuns e populares, bem como descobrir possivelmente padrões inesperados.

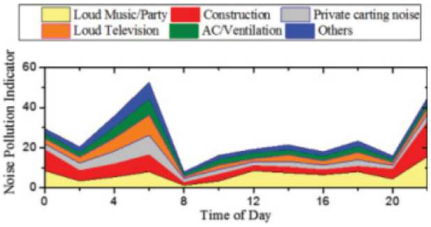
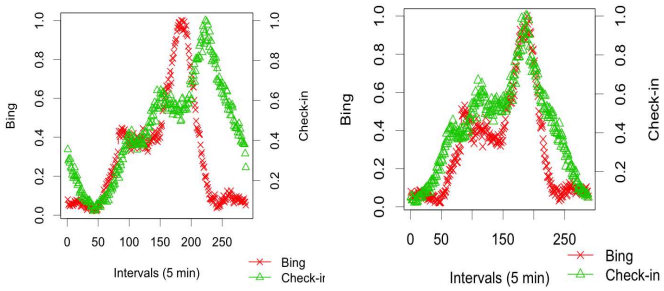


Figura 2.25. Ruídos de diferente categorias na Times Square [Zheng et al. 2014b].



(a) Intervalo de 5 minutos (b) Intervalo de 5 minutos com *shift*

Figura 2.26. Frequência de trânsito intenso em diferentes intervalos de tempo em um dia típico (segunda–sexta) [Tostes et al. 2014].

2.6.5. Problemas das Cidades

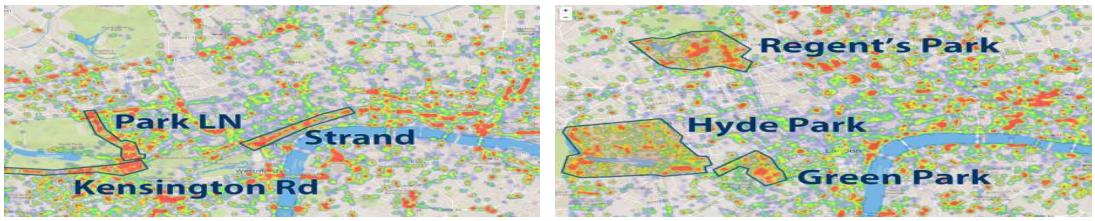
A coleta de dados sobre problemas que as cidades enfrentam pode ser facilitada com o uso de RSPs como o Colab.re. Essa RSP permite aos usuários criar, visualizar e compartilhar problemas de diversas naturezas sobre a cidade. Além desse exemplo, existem outras RSPs para o monitoramento de questões específicas do meio ambiente urbano, como o nível de ruído. Por exemplo, NoiseTube [Maisonneuve et al. 2009], como apresentamos anteriormente.

Com base no NoiseTube, D'Hondt e Stevens [D'Hondt et al. 2013] conduziram um experimento para mapear os níveis de ruído em Antwerp, Bélgica. Um dos objetivos era avaliar a qualidade dos mapas de ruído obtidos por sensoriamento participativo, em comparação com os mapas de ruído oficiais baseados em simulação. Para isso, foram realizados vários experimentos de calibração, investigando diversos aspectos dos padrões de ruído. Os autores foram capazes de construir mapas de ruído com uma margem de erro de 5dB, o que é comparado com mapas de ruído oficiais baseados em simulação.

Além dessas iniciativas, desde 2001, Nova Iorque disponibilizou uma plataforma chamada 311 para permitir que as pessoas reclamem de problemas da cidade usando um aplicativo móvel. Cada reclamação está associada a um local, data e hora, e, em alguns casos, informação detalhada da reclamação, como música alta ou barulho de construção (para os problemas de ruído). Usando os dados do serviço 311, em [Zheng et al. 2014b] os autores inferem a situação de ruído (consistindo de um indicador de poluição sonora), em diferentes momentos do dia para cada região de Nova Iorque. De acordo com o indicador de poluição sonora, é possível verificar a composição de ruído de um determinado local mudando ao longo do tempo (por exemplo, Time Square), como mostrado na Figura 2.25.

Os autores modelaram a situação de ruído de Nova Iorque com um tensor tridimensional, em que as três dimensões representam regiões, categorias de ruído e intervalos de tempo, o que permite recuperar a situação do ruído em toda a cidade. A informação de ruído não só pode facilitar a qualidade de vida de um indivíduo (por exemplo, ajudar a encontrar um lugar tranquilo para se estabelecer), mas também auxiliar os responsáveis governamentais no combate à poluição sonora.

Focados nos problemas de trânsito da cidade, em [Tostes et al. 2014] os autores estu-



(a) Natureza (b) Emissões gasosas

Figura 2.27. Mapas de cheiros em Londres [Quercia et al. 2015].

daram a seguinte pergunta: é possível utilizar dados de RSPs como uma característica para predição de trânsito intenso? Os autores observaram que os dados de RSP, especialmente os provenientes do Instagram e Foursquare, são surpreendentemente bastante correlacionados com trânsito intenso e que podem ser utilizados para desenvolver modelos de previsão de congestionamento mais eficientes. Para tratar a questão, os autores desenvolveram uma Web *crawler* para a coleta de condições de tráfego do Google Maps e do Bing Maps. De posse desses dados, bem como com dados de RSPs, os autores estudaram a cidade de Nova Iorque, mais especificamente a área de Manhattan. Conforme o estudo demonstra, os dados de RSP são bastante correlacionados com trânsito intenso, como mostrados na Figura 2.26. Ao comparar as Figuras 2.26a e 2.26b, podemos ver que a distribuição de trânsito intenso e a distribuição de *check-ins* durante os dias de semana são bastante semelhantes, no entanto as curvas são deslocadas no eixo x por um valor que pode ser calculado de acordo com uma equação proposta no estudo. Essa descoberta é surpreendente e sugere que dados de RSPs podem refletir as condições reais de tráfego. Em [Tostes et al. 2013] os autores propuseram também um modelo de previsão de tráfego utilizando uma regressão logística [Freedman 2009].

MacKerron and Mourato [MacKerron and Mourato 2013] estudaram como o ambiente local afeta a felicidade das pessoas. Para realizar esse estudo, os autores utilizaram uma aplicação para *smartphones* que tinha a finalidade de permitir aos usuários informarem o seu grau de humor, localização GPS e o nível de ruído do ambiente. Analisando mais de 3 milhões de registros de 45.000 pessoas no Reino Unido, os autores constataram que, em média, os participantes são significativamente mais felizes em ambientes externos, em contato com a natureza, do que em ambientes urbanos.

Em [Quercia et al. 2015] os autores exploraram a possibilidade de utilizar dados compartilhados em RSPs para mapear os cheiros percebidos em diversas regiões da cidade. Os resultados encontrados são promissores e mostram que essa pode ser uma nova forma para classificar áreas de acordo com o seu cheiro mais característico. Para realizar esse estudo, os autores consideraram dados do Instagram, Flickr e Twitter. Eles combinaram *tags* e *tweets* com as palavras de um "dicionário de cheiro" já existente. Em seguida, analisaram essas ocorrências na cidade. A Figura 2.27 mostra, utilizando um *heatmap*, a ocorrência de dados referentes a cheiros de natureza e emissão gasosa em Londres, Figuras 2.27a e 2.27b, respectivamente. É possível notar que o cheio de natureza é fortemente observado em parques, e o cheiro de emissão gasosa em ruas com tráfego intenso.

2.6.6. Discussão

As RSPs oferecem informações atualizadas sobre locais, bem como opiniões e preferências de seus usuários. Além disso, elas têm o potencial de tratar as questões acima mencionadas em tempo (quase) real, atingindo um elevado número de regiões do globo. Nesta seção mostramos vários estudos que servem como exemplos de como trabalhar com dados de RSPs. As informações obtidas por esses estudos podem ser úteis para o desenvolvimento de serviços e aplicações mais inteligentes. Por exemplo, entender o padrão de comportamento em determinados locais na cidade, bem como a identificação de comportamentos fora do padrão esperado pode ser muito útil para o planejamento de carga de uma rede celular urbana. Estudos que visam oferecer soluções para desafogar a transmissão de dados móveis (*mobile data offloading*) podem ter grandes benefícios ao utilizar essas informações como uma ferramenta para diminuir surpresas em demandas atuais, bem como novas demandas que podem surgir, já que a cidade está em constante mudanças. Várias outras oportunidades, bem como os desafios associados a elas, são discutidas na Seção 2.8.

2.7. Compilação de Técnicas e Ferramentas

Nesta seção discutimos as principais técnicas utilizadas nos trabalhos exemplificados nas seções anteriores, bem como algumas das tecnologias e ferramentas comumente utilizadas para a análise de dados urbanos. O objetivo não é fazer uma revisão completa da literatura. No entanto, acreditamos que os apontadores mostrados aqui podem ser bastante úteis no desenvolvimento de novas aplicações e serviços na área de computação urbana.

2.7.1. Algoritmos de Agrupamento

A realização de agrupamentos pode ser muito útil para a análise de dados urbanos, permitindo oferecer novos serviços por meio de agrupamento de características espaciais. A seguir, apresentaremos os principais algoritmos de agrupamento mencionados anteriormente.

Agrupamento baseado em densidade. Um exemplo de algoritmo popular dessa classe é o DBSCAN (*Density-based spatial clustering of applications with noise*) [Ester et al. 1996]. Algumas definições: a densidade é referente ao número de pontos dentro de um raio específico (ε); Um *core point* tem um número mínimo de pontos especificados pelo usuário ($minPts$) dentro do raio (ε); Um *border point* fica localizado na vizinhança de um *core point*; Um *noise point* é qualquer ponto que não se classifica como *core point* nem como *border point*.

A Figura 2.28 ilustra as definições mencionadas do DBSCAN, onde $minPts = 3$. O ponto A é um *core point*, pois pelo menos três pontos são vizinhos dele em um raio ε. Existem vários outros *core points* e como todos são todos acessíveis a partir de um outro *core point*, eles formam um único *cluster*. Os pontos B e C não são *core points*, mas são chamados de *border points*, porém são acessíveis a partir de A (que é um *core point*) e, portanto, também pertencem ao *cluster*. O ponto N é um *noise point*, pois não é um *core point* nem um *border point*.

Ideia do algoritmo: selecionar arbitrariamente um ponto p. Identificar todos os pontos densamente conectados a p com relação aos parâmetros ε e $minPts$. Se p é um *core point*, um *cluster* é formado. Se p é um *border point* e não há pontos densamente conecta-

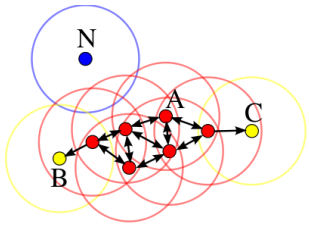


Figura 2.28. Diagrama ilustrativo do DBSCAN[25].

dos a p o DBSCAN visita o próximo ponto do conjunto de dados. O processo continua até que todos os pontos tenham sido avaliados.

Agrupamento baseado em particionamento. Algoritmos de agrupamento desta classe constroem várias partições e as avalia usando algum critério. As partições são criadas a partir da segmentação de um conjunto de dados em um conjunto de k *clusters*. O objetivo é encontrar uma partição de k *clusters* que otimiza o critério de particionamento escolhido.

O $k-means$ [Hartigan and Wong 1979] é um algoritmo bastante popular para agrupamento baseado em particionamento. Esse algoritmo divide um conjunto de N amostras em k *clusters* disjuntos. Em linhas gerais, o algoritmo possui três passos. O primeiro é escolher os centroides iniciais, por exemplo, selecionando k amostras aleatórias do conjunto N. Após a inicialização, o $k-means$ considera um laço de repetição contendo dois outros passos: atribuir cada amostra a um centroide mais próximo; e criar novos centroides calculando o valor médio de todas as amostras designadas a cada centroide anterior. Essas duas etapas se repetem até que o centroide não se mova significativamente.

Agrupamento hierárquico. O agrupamento hierárquico é um método que visa criar uma hierarquia de *clusters*. Estratégias de agrupamento hierárquico geralmente caem em dois tipos: (i) Aglomerativos: esta é uma abordagem "bottom up", ou seja, cada observação começa em seu próprio *cluster*, e a cada passo combina-se *clusters* com alguma característica comum: e (ii) Divisivos: esta é uma abordagem "top down", ou seja, todas as observações começam em um *cluster*, e divisões são realizadas de forma recursiva quando se desce na hierarquia, até que cada *cluster* tenha somente registros semelhantes.

A fim de decidir quais *clusters* devem ser combinados (por métodos aglomerativos), ou quando um *cluster* deve ser dividido (por métodos divisivos), é necessária uma medida de dissimilaridade entre os conjuntos de observações. Na maior parte dos métodos de agrupamento hierárquico, isto é alcançado pelo uso de uma medida apropriada de distância entre pares de observação e um critério de ligação (*linkage criteria*). Esse critério de ligação especifica a dissimilaridade dos conjuntos em função das distâncias entre pares de observações nos conjuntos. Um exemplo de critério de ligação é o *complete-linkage* [Sørensen 1948, Kaufman and Rousseeuw 2009]: $D_{AB} = \max\{\, d(a,b) : a \in A, b \in B \,\}$, onde d é a métrica de distância escolhida, por exemplo, distância euclidiana, e A e B são conjuntos de observações. Por exemplo, valores de D_{AB} abaixo de um limite predefinido resultariam na aglomeração de A e B.

Agrupamento espectral. Em linhas gerais, o agrupamento espectral usa a similari-

[25]https://en.wikipedia.org/wiki/File:DBSCAN-Illustration.svg.

dade entre os dados para definir os *clusters*. Para isso dois objetos matemáticos são usados: grafos de similaridade e grafos laplacianos (*graph Laplacians*). No grafo de similaridade cada vértice v_i do grafo representa um dado x_i do conjunto de dados, e dois vértices são conectados se a similaridade s_{ij} entre os dados correspondentes é positiva ou maior do que um certo limiar, com s_{ij} sendo o peso das arestas. A ideia é achar partições no grafo tal que arestas de diferentes *clusters* tenha peso baixo, enquanto que arestas dentro do mesmo *cluster* tenham pesos elevados [Luxburg 2007].

Os grafos laplacianos são as principais ferramentas para o agrupamento espectral. Diferentes algoritmos espectrais existem, cada um utiliza um grafo laplaciano descrito por von Luxburg. A ideia principal por trás do algoritmo espectral é aumentar as propriedades de agrupamento dos dados em uma nova representação de maneira que os *clusters* sejam trivialmente detectados utilizando, por exemplo, algoritmos de como o $k-means$, como nos exemplos de [Luxburg 2007].

Outros métodos de agrupamento, bem como mais detalhes sobre as técnicas de agrupamento mencionadas aqui podem ser encontrados nos livros: [Zaki and Meira Jr 2014, Kaufman and Rousseeuw 2009].

2.7.2. Regressões

Em estatística, regressão é uma técnica que permite explorar e inferir a relação de uma variável dependente (variável de resposta) com variáveis independentes específicas (variáveis explanatórias). A regressão linear foi o primeiro tipo de regressão a ser estudado rigorosamente, bem como ser usado extensivamente em aplicações práticas, incluindo a predição de valores. A regressão linear é uma equação para se estimar a condicional (valor esperado) de uma variável y, dados os valores de algumas outras variáveis x. Com isso, em geral, a regressão linear trata da questão de se estimar um valor condicional não esperado [Yan 2009].

Para se estimar o valor esperado usa-se a equação: $Y_i = \alpha + \beta X_i + \varepsilon_i$, sendo que Y_i é a variável explicada; α é uma constante, que representa a interceptação da reta com o eixo vertical; β é outra constante, que representa o declive (coeficiente angular) da reta; X_i é uma variável explicativa (independente), que representa o fator explicativo na equação; ε_i é a variável que inclui todos os fatores residuais mais os possíveis erros de medição [Yan 2009].

Outra técnica utilizada nos trabalhos citados neste minicurso é a regressão logística [Freedman 2009]. A regressão logística é um tipo de regressão que tem como objetivo produzir, a partir de um conjunto de observações, um modelo que permita a predição de valores tomados por uma variável categórica a partir de uma série de variáveis explicativas contínuas e/ou binárias.

Em comparação com as técnicas conhecidas em regressão, em especial a regressão linear, a regressão logística distingue-se principalmente pelo fato da variável resposta ser categórica. Mais detalhes sobre outros modelos de regressão, bem como os que mencionamos aqui podem ser encontrados nos livros [Harrell 2013, Chatterjee and Hadi 2015, Freedman 2009].

2.7.3. Modelos Probabilísticos para Tópicos

Modelos probabilísticos para tópicos são usados, tipicamente, para entender e estruturar grandes coleções de documentos. O modelo de tópico mais comum é o modelo de alocação latente de Dirichlet (LDA – *Latent Dirichlet Allocation*), que utiliza a abordagem bayesiana para aprender a estrutura latente de temas que compreendem cada um dos documentos. Esse modelo vem sendo utilizado em muitos trabalhos de novas áreas, incluindo a computação urbana.

O LDA se baseia em um modelo generativo assumindo que cada um dos documentos em particular na coleção é uma mistura de temas. Num modelo LDA típico, um documento de texto é representado como um conjunto de palavras, onde cada palavra pertence a um ou mais tópicos ocultos. Assim, cada documento pode ser descrito ao considerar o quanto as palavras dentro dele se relacionam com os vários tópicos ocultos e, cada tópico oculto pode ser descrito pelos termos que são mais fortemente associados a ele. Por exemplo, um documento sobre a abertura de um novo restaurante italiano pode conter as palavras "restaurante" e "jantar", associadas ao Tópico 1, e as palavras "pizza" e "spaghetti", associadas ao Tópico 2. Nesse caso, o LDA nos daria informações sobre a forma como o documento é relacionado aos dois tópicos e, com isso, poderíamos entender sobre o que esses dois tópicos dizem respeito considerando as palavras que estão associadas a eles. Por exemplo, o Tópico 2 é provável que seja sobre "alimentos" ou "comida italiana".

Uma característica bastante interessante do modelo LDA é o mínimo de intervenção humana requerida para sua aplicação. O modelo LDA é capaz de descobrir tópicos relacionados a documentos e estabelecer links entre documentos, mesmo quando não temos nenhuma informação anterior sobre estes tópicos. Além disso, os documentos não são obrigados a ser rotulados com tópicos ou palavras-chave na inicialização do modelo LDA.

2.7.4. Redução de Dimensionalidade

O objetivo dessa classe de técnicas é sumarizar os dados que contêm muitas variáveis por um conjunto menor de variáveis compostas derivadas a partir do conjunto original. Uma das motivações para o uso destas técnicas é a compressão dos dados, o que permite economizar espaço em disco e memória, bem como aumentar o desempenho de processos de aprendizado de máquina. Outra motivação é proporcionar uma maneira melhor de visualizar resultados que possuem um grande número de características (dimensões).

A técnica análise de componentes principais (*principal components analysis* - PCA [Jolliffe 2002]), é uma técnica estatística amplamente utilizada para a redução de dimensão não supervisionada. O objetivo é encontrar k vectores $u^{(1)}, u^{(2)}, ..., u^{(K)} \in \mathbb{R}^n$ para projetar os pontos de dados de modo a minimizar o erro de projeção. Em outras palavras, dado um conjunto de pontos de dados em um espaço dimensional, o objetivo é projetá-los em um espaço de dimensão menor, preservando o máximo possível de informações. A Figura 2.29 ilustra uma projeção ortogonal dos dados para uma dimensão menor: 2-D para 1-D. As linhas pontilhadas entre os pontos originais (fundo branco) e os pontos projetados na linha sólida (fundo preto) representam os erros de projeção que devem ser minimizados. Podemos utilizar o PCA, por exemplo, para encontrar a melhor aproximação planar de dados três dimensões, ou a melhor aproximação de doze dimensões dos dados com 10^4 dimensões.

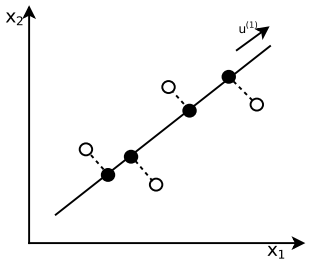


Figura 2.29. Exemplo de projeção ortogonal dos dados para uma dimensão menor: 2-D para 1-D.

Implementações para PCA podem ser encontradas em várias ferramentas. No R[26], as funções *princomp*[27] e *prcomp*[28] podem ser usadas para PCA. Em Matlab[29]/Octave[30] a função *princomp*[31] calcula os componentes principais.

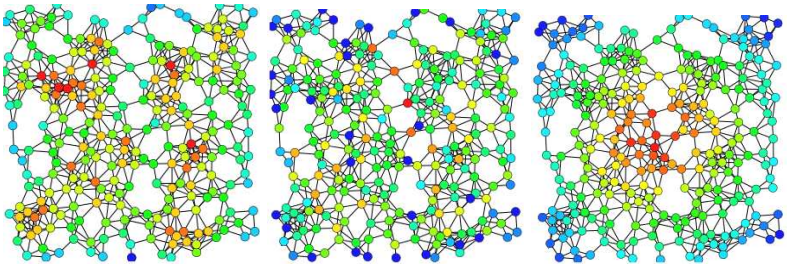
2.7.5. Análise de Sentimento

A análise de sentimento (também conhecida como mineração de opinião) refere-se ao uso de processamento de linguagem natural, análise de texto e linguística computacional para identificar e extrair informações subjetivas de materiais textuais. De um modo geral, a análise de sentimento tem como objetivo determinar a atitude de um autor com relação a algum tema ou a polaridade contextual global de um documento. A atitude pode ser o julgamento ou avaliação do autor, o seu estado afetivo (ou seja, o estado emocional quando estava escrevendo), ou a sua comunicação emocional pretendida (ou seja, o efeito emocional que se deseja ter no leitor) [Reis et al. 2015, Gonçalves et al. 2013].

A tarefa básica na análise de sentimento é classificar a polaridade da opinião expressa pelo autor em positiva, negativa ou neutra. Existem vários métodos que permitem a análise de sentimentos. No entanto, o método SentiStrength [Thelwall et al. 2010] é o exemplo aqui ilustrado, já que é bastante utilizado, possui boas taxas de precisão e foi desenvolvido para textos curtos (tamanho comumente encontrado em redes de sensoriamento participativo) [Gonçalves et al. 2013].

O SentiStrength atribui um sentimento positivo de força 1 (nenhum sentimento positivo) a 5 (sentimento positivo muito forte) e um sentimento negativo de força -1 (nenhum sentimento negativo) a -5 (sentimento negativo muito forte) para cada texto. Uma parte principal do SentiStrength é uma lista de várias palavras associadas a uma determinada força de sentimento positivo ou negativo. Por exemplo, “bom” possui valor 3 e “medo” possui valor -4, assim a frase “Eu estava com medo, mas foi bom” pode resultar 3 na escala positiva e -4 na escala negativa. Além disso, existem regras especiais para lidar com negações, perguntas, palavras de reforço (por exemplo, muito), *emoticons*, e uma série de outros casos especiais. Se for necessário um único número global para a força sentimento então o número positivo pode ser adicionado ao número negativo para dar uma pontuação no intervalo de -4 a 4

[26]http://www.r-project.org.
[27]http://stat.ethz.ch/R-manual/R-patched/library/stats/html/princomp.html.
[28]http://stat.ethz.ch/R-manual/R-patched/library/stats/html/prcomp.html.
[29]http://www.mathworks.com/products/matlab.
[30]http://www.gnu.org/software/octave.
[31]http://octave.sourceforge.net/statistics/function/princomp.html.



(a) Grau (b) Intermediação (c) Proximidade

Figura 2.30. Resultados de centralidade considerando várias métricas para o mesmo grafo[33].

[Thelwall et al. 2010].

2.7.6. Teoria de Grafos/Redes

Estruturas que podem ser representadas por grafos/redes estão em toda parte e muitos problemas de interesse prático podem ser formulados como questões sobre certos grafos [Newman 2010]. Os "grafos de transição urbana" ($G(V,E)$) representam um exemplo de grafo bastante informativo sobre a dinâmica da cidade e do comportamento social urbano. Esse tipo de particular de grafo representa, por exemplo, um conjunto V de locais na cidade (denominados vértices) e um conjunto E de pares não ordenados de V que representam a movimentação dos usuários na cidade (as chamadas arestas).

A teoria de grafos é bastante rica, oferecendo várias métricas e conceitos, como métricas de centralidade. Em particular, essas métricas determinam a importância relativa de um vértice ou aresta no grafo, que no contexto de grafos de transição urbana representam o quanto um determinado local é importante. Algumas das medidas de centralidade que são amplamente utilizadas na análise de grafos: centralidade de grau (*degree centrality*), centralidade de intermediação (*betweenness centrality*) e centralidade de proximidade (*closeness centrality*) [Newman 2010].

A centralidade de grau é definida como o número de ligações incidentes sobre um vértice, ou seja, seu grau. Em um grafo de transição urbana, um vértice com grau elevado indica um local onde as pessoas podem chegar e sair com uma alta probabilidade. Assim, a centralidade de grau é uma boa medida para identificar lugares populares da cidade. Estes locais podem ser vistos como "*hubs*" da cidade.

A centralidade de intermediação mede a importância de um vértice dentro de um grafo (existe também a intermediação das arestas, que não é discutida aqui). A centralidade de intermediação quantifica o número de vezes que um vértice age como ponte ao longo do caminho mais curto entre dois outros vértices s e t. A intermediação de um vértice v pode ser calculada assim: $C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$, onde, σ_{st} é o número total de caminhos curtos desde o vértice s ao vértice t e $\sigma_{st}(v)$ é o número desses caminhos que passam por v. Essa métrica pode ser útil, por exemplo, para indicar os locais mais interessantes para atuarem como pontes para disseminar informações entre diferentes lugares ou regiões de lugares (conjunto de lugares) na cidade.

[33]https://upload.wikimedia.org/wikipedia/commons/thumb/1/14/Centrality.svg/600px-Centrality.svg.png.

A centralidade de proximidade está relacionada com a proximidade de um vértice para todos os outros vértices na rede, ou seja, o número de arestas que separam um vértice a partir dos outros. No contexto de divulgação de informações, quanto maior a proximidade de um local, maior é a probabilidade de que a informação a ser disseminada a partir desse local alcance toda a rede na menor quantidade de tempo. Na perspectiva de um grafo de transição urbana, a centralidade de proximidade pode indicar locais estratégicos para instalar centros públicos para anúncio de informações à população. Uma ilustração do resultado de todas as métricas mencionadas aplicadas a um mesmo grafo é mostrada na Figura 2.30.

Existem várias outras métricas e conceitos de teoria de grafos além dos que foram mencionadas no texto do minicurso. Uma revisão detalhada das principais métricas e conceitos pode ser encontrada em [Newman 2010, Newman 2003, Easley and Kleinberg 2010]

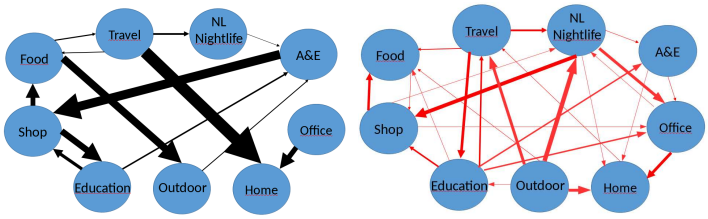
2.7.7. Validações

Dados de RSPs podem possuir várias limitações. Em primeiro lugar, eles podem refletir o comportamento dos cidadãos que utilizam as RSPs: usuários que tendem a ter menos de 50 anos, donos de *smartphones* e moradores urbanos [Brenner and Smith 2013, Duggan and Smith 2014]. Além disso, os usuários podem não compartilhar dados em todos os seus destinos, por exemplo, hospitais e motéis. Em segundo lugar, os dados podem ser baseados em uma amostra limitada de dados. Isso significa que podemos ter apenas uma amostra das atividades realizadas. Fatores externos, tais como más condições meteorológicas, podem afetar o número total de dados que são coletados para alguns lugares, especialmente locais ao ar livre. Por isso, antes de tirar conclusões com dados de RSPs, é necessário realizar uma comparação com dados obtidos de maneira tradicional (*offline*). Existem várias fontes de dados disponíveis na Web que permitem realizar validações, algumas delas foram ilustradas em [Barbosa et al. 2014].

Para investigar a precisão do método para a identificação das fronteiras culturais [Silva et al. 2014c], os autores compararam os resultados obtidos com os resultados do World Values Surveys (considerado o mais importante da área disponível publicamente), que utilizou dados coletados de forma tradicional: questionários. Além disso, em [Silva et al. 2013c] os autores utilizaram dados do TripAdvisor (um popular *website* de turismo), para verificar se os resultados dos pontos turísticos encontrados eram relevantes na cidade estudada. Outro exemplo pode ser encontrado em [Cranshaw et al. 2012], onde os autores entrevistaram moradores da cidade estudada para validar os resultados obtidos.

2.7.8. Tratamentos Estatísticos

No estudo de propriedades de grafos, um modelo nulo (*null model*) é um grafo que corresponde a um grafo específico em algumas das suas propriedades estruturais, mas que também é considerado um exemplo de um grafo aleatório. O modelo nulo é usado, geralmente, como um termo de comparação para verificar se o grafo em questão apresenta alguma característica, como a estrutura de uma comunidade, ou não. Nem todos os modelos nulos são iguais. A criação de cada um depende de quais propriedades serão preservadas no grafo original e, também, qual é a hipótese nula sendo feita. Nas técnicas City Image e identificação de POIs, mencionadas anteriormente, foram utilizados modelos nulos que mantinham os vértices de grafos de transição urbana e distribuíam o mesmo número de arestas originais de forma



(a) Original (b) Modelo nulo

Figura 2.31. Grafo de transição urbana original e um modelo nulo.

aleatória (procedimento ilustrado na Figura 2.31, onde a espessura da aresta representa o seu peso). Esses modelos serviram para identificar transições que poderiam ser geradas de forma aleatória, bem como identificar transições populares. Além disso, este procedimento é importante porque pode nos impedir de usar dados que não têm relação com os fenômenos que estamos interessados.

Outro procedimento necessário para comparar propriedades de diferentes regiões (por exemplo, cidades), é realizar uma normalização dos dados [Freedman 2009]. Além disso, também é interessante verificar a razão de dados por habitantes, valor que, idealmente, deve ser similar para todas as regiões estudadas. Existem várias outras técnicas para realizar validações e tratamento estatístico de dados de RSPs para estudar o funcionamento de cidades e o comportamento urbano. A principal mensagem aqui é ter em mente que esses procedimentos são extremamente necessários e devem ser usados sempre que possível.

2.7.9. Aterfatos para o Estudo de Mobilidade

Nesta seção, apresentamos mais detalhes sobre algumas das principais técnicas que foram utilizadas para o estudo de mobilidade em trabalhos citados anteriores.

Raio de Giro. O raio de giro mede a frequência e o quão longe um usuário se move. Um baixo raio de giro normalmente indica um usuário que viaja principalmente localmente (com poucos compartilhamentos de dados de longa distância), enquanto um alto raio de giro indica um usuário com muitos compartilhamento de dados de longa distância [Cheng et al. 2011]. O raio de giro de um usuário pode ser formalizado como:

$r_g = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\mathbf{r}_i - \mathbf{r}_{\text{cm}})^2}$, onde n é o número de dados do usuário, e $\mathbf{r}_i - \mathbf{r}_{\text{cm}}$ é a distância entre um determinado dado $\mathbf{r}_i$ e centro de massa do usuário $\mathbf{r}_{\text{cm}}$ (que é uma média simples da localização de todos os dados).

Modelo de gravidade. Na sua formulação mais simples, o modelo de gravidade (*Gravity Model* [Zipf 1946]) afirma que a interação $T_{i,j}$ entre dois locais i e j é proporcional ao produto das suas populações P_i e P_j sobre a sua distância $d_{i,j}$:

$T_{i,j} = k\frac{P_i^{\alpha}P_j^{\beta}}{d_{i,j}^{\gamma}}$, onde os expoentes α, β, γ e o fator de escala k são parâmetros ajustáveis, praticamente escolhidos de modo a ajustar os dados empíricos que estão sendo modelados. Ao longo dos anos, este modelo simples foi ampliado e refinado em uma variedade de maneiras, e sua aplicação tem ocorrido além do domínio de transportes, o que mostra o seu potencial. Por exemplo, ele foi usado para modelar telefonemas interurbanos

[Krings et al. 2009] e a propagação de doenças infecciosas [Balcan et al. 2009].

Fórmula de Haversine. A fórmula de Haversine é uma importante equação usada em navegação, fornecendo distâncias entre dois pontos de uma esfera a partir de suas latitudes e longitudes, podendo ser calculada como: $d = 2r\arcsin\left(\sqrt{\sin^2\left(\frac{\phi_2-\phi_1}{2}\right) + \cos(\phi_1)\cos(\phi_2)\sin^2\left(\frac{\lambda_2-\lambda_1}{2}\right)}\right)$, onde d é a distância entre os dois pontos (ao longo de um grande círculo, *great circle*, da esfera); r é o raio da esfera (raio da Terra = 6372,8 km); ϕ_1, ϕ_2: latitude do ponto 1 e latitude do ponto 2 e λ_1, λ_2: longitude do ponto 1 e longitude do ponto 2.

Existem várias outras técnicas e modelos que auxiliam o estudo de mobilidade urbana utilizando dados de RSPs, além dos trabalhos citados, mais informações podem ser encontradas em outras áreas que também realizam estudos neste campo, como geografia, planejamento urbano e física.

2.7.10. Tecnologias e Ferramentas para a Análise de Dados

A análise dos dados urbanos é fundamental, pois para utilizar esses dados é necessário conhecer suas propriedades. O objetivo desta seção é apresentar algumas das tecnologias e ferramentas comumente utilizadas para essa finalidade.

Quando se trata de análise de dados, a linguagem Python tanto quanto as linguagens Matlab/Octave ou R são relativamente fáceis para começar a codificar. A seguir mostramos algumas das vantagens de cada uma destas ferramentas.

Quando usar as linguagens Matlab ou R: Matlab e R possuem uma longa e confiável história, uma comunidade forte e com uso na indústria. Isto significa que se pode contar com o apoio online de outros usuários, caso precisar de ajuda ou tenha dúvidas sobre a utilização da linguagem. Além disso, há uma grande variedade de pacotes extras disponíveis publicamente. Quando é necessário fazer uma análise estatística pesada ou gráficos, Matlab ou R são as melhores opções. Operações matemáticos típicas como multiplicação de matrizes funcionam de forma bem simples e direta, bem como a visualização de dados.

Quando usar a linguagem Python: Se é necessário realizar uma limpeza e formatação dos dados, ou obter dados de *websites*, arquivos ou outras fontes de dados, a melhor escolha é Python. Em geral, devido à facilidade de transformar ideias em código e pela simplicidade de manipular *strings*, Python é uma das principais linguagens para a construção de *scripts* para o gerenciamento de dados. Como ilustramos anteriormente, existem várias APIs desenvolvidas para a linguagem Python que facilitam a coleta de dados de RSPs.

Outras ferramentas úteis na análise de dados de RSPs:

- **Estudo de grafos/redes:** Uma boa opção é o NetworkX[34]. NetworkX é um pacote para Python que permite a criação, manipulação e estudo de grafos de forma bem simples;
- **Visualização de grafos:** Para essa finalidade, duas boas opções são o Gephi[35] e o Cy-

[34]https://networkx.github.io.
[35]http://gephi.org.

toscape[36]. Ambos são grátis, e além de possibilitar a construção de boas visualizações, também oferecem suporte básico para a análise do grafo;

- **Manipulação de dados:** Como mencionamos anteriormente, muitas fontes de dados urbanos oferecem um volume muito grande de dados (*Big Data*). Com isso, o uso de bancos de dados NoSQL [Pokorny 2013], como o MongoDB[37], se tornam interessantes, pois são projetados para oferecer escalabilidade na manipulação de um volume grande de dados;

- **Análises estatísticas e visualização de dados (não espaciais) em Python:** Assim como as linguagens Matlab ou R, a linguagem Python também pode ser utilizada para realizar análises estatísticas e visualização de dados. Para que a linguagem Python tenha funcionalidades similares à linguagem Matalb ou R é necessário a instalação de pacotes específicos, por exemplo: NumPy[38], SciPy[39], Matplotlib[40] e Scikit-learn[41];

- **Processo de geocodificação (*geocoding*):** Processo de conversão de endereços (por exemplo, "1600 Amphitheatre Parkway, Mountain View, CA") em coordenadas geográficas (por exemplo, latitude 37.423021 e longitude -122.083739), que podem ser usadas para a análise espacial de dados urbanos. O Google[42], Bing Maps[43], Yahoo[44] e Mapquest[45] oferecem APIs para essa finalidade. Existem bibliotecas para linguagens de programação que auxiliam nessa tarefa, por exemplo, a biblioteca para Python Geopy[46];

- **Verificar existência de lei de potência (*Power law*)**: Como mostramos anteriormente, ao analisar dados de RSPs podemos querer verificar se uma determinada distribuição de dados pode ser explicada por uma lei de potência. Para isso, existe um conjunto de ferramentas bastante útil disponibilizado em: http://www.santafe.edu/∼aaronc/powerlaws. Mais detalhes sobre este processo, pode ser obtido também em [Alstott et al. 2014];

- **Análise e visualização de dados especiais:** Existem várias ferramentas para essa finalidade, no entanto mencionamos aqui apenas algumas delas. Uma das possibilidades é utilizar um Sistema de Informação Geográfica (GIS - *Geographic Information System* [Sutton et al. 2009]). Uma aplicação GIS, por exemplo o QGIS[47], auxilia no processo de manipulação de dados espaciais, possibilitando a visualização de informações em um mapa e a realização de análises espaciais diversas. Além disso, existem várias bibliotecas em diversas linguagens de programação que permitem a manipulação de

[36]http://www.cytoscape.org.
[37]https://www.mongodb.com.
[38]http://www.numpy.org.
[39]http://www.scipy.org.
[40]http://matplotlib.sourceforge.net.
[41]http://scikit-learn.org.
[42]https://developers.google.com/maps/documentation/geocoding/?hl=pt-br.
[43]https://msdn.microsoft.com/en-us/library/ff701715.aspx.
[44]https://developer.yahoo.com/boss/placefinder.
[45]http://www.mapquestapi.com/geocoding.
[46]https://github.com/geopy/geopy.
[47]http://www.qgis.org.

Tabela 2.1. Comparando diferente alternativas para visualizar e manipular dados espaciais.

	GIS (por exemplo, QGIS)	Python	R
Instalação e Configuração do ambiente de trabalho	Fácil instalação. Interfaces gráficas relativamente intuitivas. Disponível para os principais sistemas operacionais (SOs).	Geralmente possui dependências de pacotes na instalação. Independente de SO.	Depende da instalação de pacotes, processo que é geralmente fácil. Independente de SO.
Análise de dados	Bom para iniciantes.	Tende a ser a plataforma estatística preferida para cientistas com elevado conhecimento de programação. Oferece flexibilidade.	Tende a ser a plataforma estatística preferida para cientistas de áreas diversas que usam R. A curva de aprendizado é a mais baixa. Oferece flexibilidade.
Criação de um mapa	Opção mais fácil.	Requer vontade de aprender fundamentos de programação.	Bom para cientistas que já usam R.

dados espaciais. Por exemplo, o pacote ggmap [Kahle and Wickham 2013] baseado em R disponibiliza um conjunto de funções para visualizar dados e modelos espaciais em cima de mapas estáticos a partir de várias fontes online (por exemplo, Google Maps). Várias bibliotecas da linguagem Python também oferecem suporte para essa tarefa. Por exemplo, a biblioteca Cartopy[48] facilita a visualização de dados espaciais em mapas estáticos, já a bilioteca Folium facilita a visualização de dados espaciais um mapa interativo, com o suporte da biblioteca Leaflet[49]. Especificamente sobre *heatmaps*, a biblioteca Python Heatmaps (jjguy)[50], oferece *heatmaps* compatíveis com Google Earth e a Python Heatmaps (sethoscope)[51], oferece *heatmaps* baseados no OpenStreetMap[52]. Vários outras bibliotecas baseadas em Python para trabalhar com dados espaciais pode ser encontrada aqui: [Jung 2015]. Como temos muitas opções para realizar as mesmas tarefas, a Tabela 2.1 visa apresentar mais algumas características das alternativas mencionadas com o intuito auxiliar numa eventual escolha de uso.

2.8. Desafios

Esta seção apresenta diversos desafios sobre tópicos de pesquisa atuais relacionados com computação urbana utilizando dados de redes de sensoriamento participativo.

[48]http://scitools.org.uk/cartopy.
[49]http://leafletjs.com.
[50]http://jjguy.com/heatmap.
[51]http://www.sethoscope.net/heatmap.
[52]https://www.openstreetmap.org.

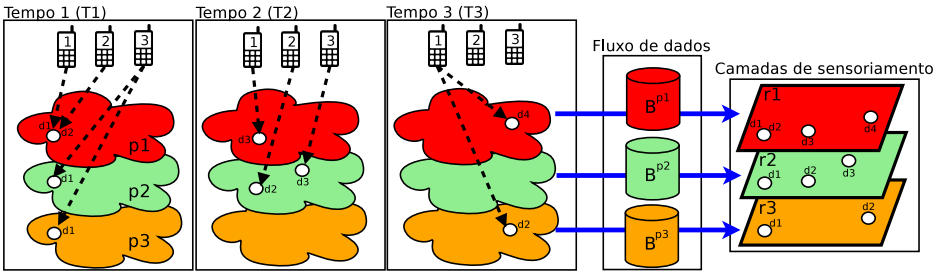


Figura 2.32. Ilustração do compartilhamento de dados em três RSPs ao longo do tempo, resultando em camadas de sensoriamento [Silva et al. 2014b].

2.8.1. Camadas de Sensoriamento

Uma camada de sensoriamento consiste de dados descrevendo aspectos específicos de uma localização geográfica. O conceito de camada de sensoriamento é bastante amplo: ele representa dados, com seus atributos, provenientes de uma determinada fonte de dados, por exemplo uma RSP particular. Cada RSP fornece acesso aos dados relacionados a certo aspecto de uma região geográfica predefinida (por exemplo, condições de tráfego, fotos de locais, etc.), e, com isso, cada RSP distinta pode ser representada como uma camada de sensoriamento [Silva et al. 2014b].

Além de RSPs, outras fontes de dados podem ser: dados disponíveis na Web não gerados por usuários, como a condição climática fornecida pela empresa Clima Tempo[53]; ou dados de redes de sensores sem fio tradicionais. Discutimos aqui o conceito de camadas de sensoriamento para as RSPs. No entanto todos os conceitos discutidos podem ser utilizados para outras fontes de dados associadas a regiões geográficas predefinidas, com as adaptações necessárias.

A Figura 2.32 ilustra o conceito de camadas de sensoriamento. Essa figura mostra dados compartilhados em três RSPs diferentes (p1, p2 e p3), por três usuários distintos em diferentes instantes de tempo. Como discutimos na Seção 2.3, esses dados devem ser coletados (por exemplo, usando uma API) e processados, passo que inclui as tarefas de análise e padronização dos dados. Cada plano na figura representa uma camada de sensoriamento para uma região específica, por exemplo o centro de Belo Horizonte, com dados provenientes de três fontes distintas. Com isso, as camadas de sensoriamento ilustradas são: *check-ins* (r1), proveniente, por exemplo, do Foursquare; *alertas de tráfego* (r2), proveniente, por exemplo, do Waze; e *fotos de lugares* (r3), proveniente do Instagram, por exemplo.

O uso de camadas de sensoriamento de forma independente pode ser muito útil. Por exemplo, uma camada contendo informações de trânsito pode possibilitar a identificação em tempo real de rodovias com acidentes e buracos, cuja detecção é difícil com sensores tradicionais, mas torna-se mais factível quando os usuários participam do processo de sensoriamento. Os exemplos mencionados na Seção 2.6 também contribuem para esse ponto. No entanto, a grande motivação é realizar uma análise conjunta de múltiplas camadas para a construção de aplicações mais sofisticadas.

Sabemos que uma queixa comum dos habitantes das grandes cidades é o congestio-

[53] http://www.climatempo.com.br.

namento. Com isso, uma aplicação que emerge naturalmente é uma que possui o objetivo de inferir as causas de congestionamento, passo fundamental para tratar o problema. Esta não é uma tarefa fácil de realizar, e o resultado pode variar entre diferentes cidades, uma vez que dependem de aspectos geográficos, culturais, econômicos, dentre outros. No entanto, a análise conjunta de diferentes camadas de sensoriamento na cidade poderia contribuir para essa aplicação. Por exemplo, poderíamos cruzar as informações fornecidas pelas camadas alertas de tráfego, *check-ins* e fotos de lugares. A primeira camada fornece dados em tempo (quase) real sobre onde estão acontecendo congestionamentos, a segunda fornece dados sobre os tipos de lugares localizados nas áreas dos congestionamentos, com isso é possível entender melhor as áreas de interesse, por exemplo, identificando o tipo da área e, finalmente, através da análise da camada fotos de lugares nós podemos obter evidência visual do que acontece em tempo real próximo das áreas durante os congestionamentos. Ao analisar conjuntamente dados destas três camadas podemos detectar, por exemplo, carros bloqueando cruzamentos, e inferir as possíveis causas disso. Obviamente, outras camadas podem também ser utilizadas, tal como a condição do clima.

Há vários desafios ao lidar com dados de várias camadas simultaneamente, como os descritos a seguir.

1. **Combinação de dados**: A fim de realizar a combinação de dados nós temos que certificar que estes são consistentes em todas as camadas. Esta é uma condição obrigatória para a correta extração de informações. Por exemplo, para combinar dados compartilhados pelo mesmo usuário em diferentes camadas pode ser um problema em RSPs, pois o mesmo usuário pode participar em diferentes camadas com diferentes identificadores. Vamos supor que queremos combinar dados de um mesmo usuário compartilhados na camada *check-ins* e na camada fotos de lugares. Como os dados dessas camadas são de sistemas independentes os usuários possuem identificações diferentes. Uma forma de tentar contornar esse problema é verificar outros sistemas com o intuito de mapear o ID do usuário de uma camada em outra. Sabemos, por exemplo, que os usuários do Foursquare e Instagram tendem a ser também usuários do Twitter. Dessa forma, a chave do processo de combinação poderia ser a identificação usada no Twitter.

   Além disso, os sistemas de computação urbana são, tipicamente, obrigados a responder rapidamente a consultas dos usuários (por exemplo, para prever condições de tráfego). Sem técnicas de gerenciamento de dados que permitem organizar várias camadas de sensoriamento heterogêneas, torna-se impossível realizar processos de extração de conhecimento de forma rápida. Esses são apenas alguns dos desafios relacionados a esse tópico.

2. **Validade dos dados:** Diferentes camadas podem se referir a dados válidos para diferentes intervalos de tempo. Isso é natural porque algumas fontes de dados fornecem dados em tempo (quase) real, outras não. Por exemplo, um alerta no Waze refere-se a uma situação de trânsito que pode não existir cinco minutos mais tarde. No entanto, dados do censo geralmente são válidos por um grande intervalo de tempo, meses ou anos, até o próximo censo ser publicado. Temos de estar cientes de todas essas questões ao projetar novas aplicações.

2.8.2. Dinâmica Temporal

Um aspecto pouco explorado nas análises dos dados de RSPs é o temporal. Nesse contexto, a análise de características temporais permite aprimorar tais aplicações, bem como gerar novas oportunidades de pesquisa [Gao et al. 2013, Yuan et al. 2013]. A maioria dos estudos encontrados na literatura consideram que os dados compartilhados por usuários formam RSPs estáticas, sendo a dinâmica temporal negligenciada. Isso pode acarretar em perda de informações importantes. Por exemplo, enquanto duas regiões de uma cidade podem apresentar comportamento similar nos dados agregados durante um dia, elas podem ter diferenças quando uma perspectiva temporal é considerada na análise das atividades mais populares em cada turno.

Para ilustrar algumas iniciativas nesse sentido, em [Zhang et al. 2013] os autores analisaram atividades urbanas a partir de dados do Foursquare considerando a dinâmica temporal, a partir da divisão dos dados em períodos do dia. Os autores identificaram, por exemplo, que a atividade "comida" pode não ser tão ativa no período da tarde, mas sim, nos períodos da manhã e noite, de forma que na visão agregada não se perceba essa diferença. Essa abordagem é interessante para mostrar que determinadas atividades são pertinentes em um determinado período do dia, mas quando analisadas de forma agregada podem não ser relevantes ou podem não capturar o real comportamento dos usuários. Outro exemplo é a técnica City Image que apresentamos na Seção 2.6.1. No exemplo da Figura 2.19, a perspectiva temporal utilizada é a divisão dos dados em dias de semana/final de semana durante o dia e a noite, a partir disso foi feita uma análise dos dados por partições. Podemos perceber com o auxílio dessas imagens que existe variação significativa entre dia e noite nas duas cidades analisadas. Além disso, a imagem agregada (sem considerar partições) é bastante diferente das desagregadas, como foi mostrado em [Silva et al. 2012].

Os trabalhos descritos anteriormente fornecem indícios das vantagens da utilização de informação temporal dos dados obtidos de RSPs. No entanto, se por um lado investigar a dinâmica temporal de uma RSP é uma oportunidade para obtenção de informações mais próximas da realidade do comportamento da rede, por outro, surgem novos desafios ao adicionarmos uma dimensão temporal ao estudo, como os descritos a seguir:

1. **Janelas de tempo:** Trabalhos que analisam a questão temporal geralmente fragmentam os dados em intervalos de tempo (e.g., manhã, tarde e noite) denominados janelas. No entanto, a definição adequada do tamanho da janela é um problema, pois é necessário definir um tamanho de janela que capture dinâmicas relevantes. Nesse caso, existem inúmeras oportunidades para novas abordagens que consideram janelas com tamanhos flexíveis;

2. **Modelagem:** Geralmente, dados de uma RSP são representados como um conjunto de entidades, por exemplo, usuários ou PDIs (pontos de interesses), e suas relações (e.g., transições ou comunicação). Como a contribuição desses dados pode variar muito ao longo do tempo um modelo baseado em grafos estáticos pode não ser suficiente para capturar essa dinamicidade. Por exemplo, dados obtidos a partir do Foursquare possuem informações espaço-temporais, tais como posicionamento dos usuários e os momentos de interação. Portanto, um desafio é modelar a dinâmica espaço-temporal a fim de entender melhor, por exemplo, diversos aspectos dessa participação dos usuá-

rios. Nesse sentido, grafos temporais [Kostakos 2009] surgem como alternativa promissora que pode ser utilizada para entendimento da dinâmica espaço-temporal. Em um grafo temporal, as relações entre as entidades podem ser modeladas como arestas que podem ser criadas e destruídas ao longo do tempo. Por exemplo, entender aspectos temporais de interações entre usuários com certos locais na cidade. Dessa forma, utilizando grafos temporais para modelar uma RSP, podemos aplicar tanto conceitos de grafos (e.g., componentes conectados e caminhamento) como métricas de centralidade (e.g., centralidade de intermediação e centralidade de proximidade) para auxiliar no entendimento da dinâmica das RSPs [Nicosia et al. 2013].

2.8.3. Mecanismos de Incentivos para a Aquisição de Dados

Um ponto fundamental para as RSPs é a colaboração dos usuários, pois as aplicações em uma RSP dependem de que os usuários estejam dispostos a coletarem, processarem e transmitirem os dados sensoriados [Lee and Hoh 2010]. A colaboração entre os participantes de uma RSP reflete diretamente na qualidade e quantidade dos dados sensoriados e, consequentemente, na melhoria dos serviços oferecidos com esses dados.

No entanto, como estas aplicações consomem recursos do dispositivo do usuário, o mesmo pode ser relutante em contribuir com a rede. Diversos são os motivos que podem fazer um usuário usufruir, porém não colaborar com a RSP, tais como poupar bateria, evitar gastos com a transmissão de dados, ou mesmo por questões de privacidade [Lee and Hoh 2010].

Nos últimos anos, foram propostos dezenas de mecanismos de incentivo e realizados diversos experimentos para entender o comportamento cooperativo. Alguns desses mecanismos podem recompensar o participante por meio de pagamentos reais, virtuais ou outros prêmios. Outros mecanismos visam transformar a tarefa de sensoriamento em uma tarefa mais prazerosa e estimulante para o usuário, adicionando elementos comuns em jogos, como elementos de disputa entre os usuários (mecanismos conhecidos como baseados em *gamificação*).

Existem mecanismos que visam permitir que o usuário participe da decisão sobre a tarefa que irá realizar e sobre o pagamento que irá receber da RSP. Outros visam melhorar a qualidade dos dados obtidos e minimizar os custos com sensoriamento. Podemos mencionar também mecanismos em que o usuário negocia com a plataforma o valor da recompensa pelos dados sensoriados antes de enviá-los e, ainda, os que a plataforma decide quanto irá pagar pelos dados já enviados pelo usuário.

Um mecanismo de incentivo é eficiente se ele recruta mais participantes para a RSP e mantém esses participantes ativos no sistema. Com isso alguns dos desafios relacionados aos mecanismos de incentivos para as RSPs são:

1. **Custos**: Para que o desenvolvimento de mecanismos de incentivo monetários sejam eficientes, deve-se considerar os custos para a plataforma da RSP e os ganhos para o participante da rede. Esses mecanismos utilizam um custo máximo para a plataforma RSP que será pago aos participantes ativos da rede. No entanto, encontrar e decidir um valor que minimize o custo para plataforma e, ao mesmo tempo, motive o usuário requer investigações futuras [Gao et al. 2015].

2. **Validação das propostas**: A maioria das propostas de mecanismos de incentivo utilizam uma validação teórica ou pequenos experimentos controlados. No entanto, estes experimentos podem não predizer com alta precisão a participação dos usuários ao longo do tempo na plataforma. No caso de mecanismos baseados em *gamificação*, embora diversas RSPs de sucesso no mercado utilizem este conceito, aplicar uma estratégia com sucesso prévio em uma nova RSP não é garantia que funcionará. Talvez exista alguns elementos que funcionem para determinados tipos de RSPs e outros não.

2.8.4. Qualidade de Dados

A qualidade de dados é um tópico amplamente estudado pela comunidade científica. No entanto, ainda existem desafios únicos para controlar a qualidade dos dados compartilhados quando se lida com contribuições de usuários ubíquos.

Em geral, como discutimos anteriormente, os dados coletados de RSPs são, após processados, utilizados para a extração de informações contextuais, que são fundamentais para os sistemas sensíveis ao contexto [Dey and Abowd 2000].

Alguns dos principais desafios relacionados à qualidade dos dados em uma RSP são:

1. **Erros de leitura**: Um desafio que pode afetar a precisão dos dados das RSPs são possíveis erros de leitura de equipamentos. Por exemplo, um GPS pode estar mal calibrado e gerar dados cuja imprecisão está além do limite aceitável para este tipo de dado. Embora alguns erros possam parecer totalmente toleráveis, dependendo dos requisitos de uma aplicação, é possível que os limites mais restritos de precisão sejam fundamentais para sua correta operação.

2. **Ausência de estrutura**: Os dados compartilhados em RSPs, em alguns casos, são de texto livre, não apresentando uma estrutura semântica nem codificadas. Essa liberdade dada aos usuários permite que eles postem o que querem, mesmo informações incorretas, e em diferentes formatos. Por exemplo, um usuário poderia descrever um acidente em outra língua ou utilizando gírias através de algum *microblogging* como o Twitter. Com isso, o processamento dos dados se torna complexo e suscetível a erros, uma vez que há a possibilidade de dados distintos serem confundidos como um mesmo dado, ou ainda a duplicidade de dados, isto é, dados idênticos serem identificados como distintos devido a diferenças no preenchimento dos campos.

3. **Poluição dos dados**: Este desafio diz respeito à possibilidade dos dados estarem incorretos devido a um comportamento malicioso dos usuários
[Coen-Porisini and Sicari 2012, Mashhadi and Capra 2011]. Podemos encontrar comportamentos maliciosos em várias esferas sociais, e o mesmo também pode ocorrer nas RSPs. Por exemplo, usuários de sistemas para compartilhamento de alertas de trânsito, como o Waze, podem gerar falsos alertas de congestionamento ou acidentes, com o intuito de incentivar os demais usuários a não utilizar determinadas vias de seu trajeto. Este comportamento malicioso poderá ocasionar em falsos positivos, por exemplo, na detecção de eventos.

2.8.5. Outros Desafios

Outra questão importante é lidar com um grande volume de dados que as RSPs podem oferecer, impondo desafios para armazenamento, processamento e indexação em tempo real usando ferramentas de gerenciamento de banco de dados tradicionais ou aplicações de processamento de dados. Isso faz com que a oferta de serviços em tempo real usando uma rede de sensoriamento participativo seja um desafio. Para resolver esta questão, precisamos de métodos para armazenar, mover e processar de forma eficaz grandes quantidades de dados. Novos paradigmas algorítmicos devem ser projetados, bem como técnicas de mineração de dados específicas devem ser criadas de acordo com esses novos paradigmas. Outros métodos devem contemplar abordagens de engenharia de dados para grandes redes com milhões ou bilhões de nós/arestas, incluindo compressão eficaz, pesquisa e métodos para casamento de padrões [Giannotti et al. 2012].

Felizmente, a pesquisa sobre os desafios de grandes dados é muito ativa, e recentemente fez grandes avanços, por exemplo, com base em plataformas paralelas (como o Hadoop[54]), para o processamento de um grande volume de dados.

Podemos ainda mencionar que as RSPs são muito dinâmicas. Usuários contam com RSPs, como o Twitter ou Waze, para transmitir seus dados sensoriados. De posse desses dados podemos extrair conhecimento. Sistemas, como o Waze, por sua vez, podem ser realimentados com esses conhecimentos obtidos e, a partir disso, eles podem fornecer informações úteis para os usuários. Esses conhecimentos também podem ser gerados por aplicativos de terceiros. Para exemplificar, na Seção 2.6 é descrito um exemplo de aplicação que permite a identificação de regiões de interesse de uma cidade. Após usar este aplicativo, os usuários podem optar por mudar o seu comportamento, como para visitar preferencialmente áreas populares, o que pode vir a afetar o número de dados compartilhados nesses locais. Isto dá uma ideia do dinamismo de uma rede de sensoriamento participativo e os desafios que surgem nessas condições.

Além desses desafios, existe ainda a questão da privacidade do usuário. Essa questão é bastante ampla e está presente em muitas camadas do sistema. Privacidade de dados em sistemas de mídia social atualmente tem sido discutida em vários estudos, como em [Pontes et al. 2012, Toch et al. 2010, Brush et al. 2010].

2.9. Oportunidades

Nesta seção ilustramos algumas das oportunidades relacionadas com o estudo de sociedades urbanas utilizando dados de RSPs.

Detecção/previsão de trânsito: De forma geral, dados de RSPs são pouco explorados em modelos de detecção/previsão de trânsito. Alguns dos trabalhos mais próximos dessa direção são: [Silva et al. 2013e, Tostes et al. 2014]. Tostes et al. analisaram condições de trânsito em relação aos dados sensoriados por duas RSPs, Foursquare e Instagram. Como vimos na Seção 2.5 os sensores de RSPs fornecem informações importantes para o melhor entendimento da dinâmica das cidades. Por exemplo, uma mensagem geolocalizada, seja no Foursquare, Instagram ou Twitter, pode ser utilizada para melhor entendermos as condições de trânsito, como foi mostrado em [Tostes et al. 2014] (trabalho discutido na Se-

[54]http://hadoop.apache.org.

ção 2.6.5). Além disso, imagine que um usuário faça um *check-in* em casa e depois vá para o trabalho. Quando chegar no trabalho, por algum motivo, ele faz outro *check-in*. Independente se for na mesma rede social ou não, existe uma informação intrínseca no intervalo de tempo entre esses *check-ins* que consiste no desempenho do trânsito. Se o trânsito estiver mais congestionado, esse intervalo entre *check-ins* será maior do que o tempo de viagem sem congestionamento, que é facilmente calculado pela distância e velocidade máxima das vias. Além disso, os autores também levantaram várias questões nessa direção: (i) como coletar dados de mapas online em tempo real?; (ii) é possível utilizar dados de RSPs como uma característica para previsão de trânsito intenso?

Modelagem de camadas de sensoriamento: Temos ainda oportunidades com relação à modelagem das camadas de sensoriamento, pois numa mesma camada as entidades podem possuir relações distintas entre elas. Para ilustrar essa oportunidade considere a camada de *check-ins*. Como ilustramos anteriormente, essa camada pode representar a mobilidade urbana considerando a relação entre lugares e pessoas, sendo útil para entender, por exemplo, a frequência de transição entre diferentes lugares (entidades). Outra possibilidade é modificar a modelagem do problema, para, por exemplo, estudar as preferências de pessoas. Nesse caso, a entidade a ser analisada passa a ser o usuário. Note que dados de uma mesma camada podem ser modelados de formas distintas para responder perguntas diferentes. Alguns arcabouços, como o apresentado em [Silva et al. 2014b], oferecem suporte básico para essa questão. No entanto, existem várias oportunidades para desenvolver extensões desse arcabouço a fim de oferecer serviços mais sofisticados.

Múltiplas estratégias de mecanismos de incentivo: A maioria das propostas para incentivar a colaboração nas RSPs focam em apenas uma estratégia. Porém, como observado em [Reddy et al. 2010], a utilização de mais de uma estratégia simultaneamente pode apresentar melhores resultados. Esses autores concluem que os incentivos funcionaram melhor quando os pagamentos (recompensas) foram combinados com outros fatores, como o altruísmo do usuário e quando havia competição entre os participantes. Além disso, mostraram que um pagamento justo para todos os participantes os mantinham mais motivados do que micro pagamentos.

Confiabilidade de um determinado usuário: Dados gerados por usuários mais confiáveis provavelmente terão maior qualidade. Uma possível direção nesse sentido está relacionada com a identificação de padrões comportamentais dos usuários das RSPs. Como apresentado na Figura 2.10, quando grandes quantidades de dados são agregadas, é possível, claramente, identificar padrões de comportamento no compartilhamento de dados durante diferentes dias da semana. Assumindo que este conhecimento prévio seria uma referência do comportamento esperado dos usuários de uma dada RSP, uma possibilidade seria comparar o comportamento de um determinado usuário com este padrão de referência. Por exemplo, usuários que possuem um padrão de compartilhamento muito distinto dos demais poderia representar um usuário não confiável (e.g., um robô malicioso).

Essa abordagem discutida acima, pode ser caracterizada como uma espécie de filtragem colaborativa [Adomavicius and Tuzhilin 2005]. Esta é uma estratégia utilizada por sistemas de recomendação quando não se possui um conhecimento prévio do usuário ao qual deverá se recomendar um item. Por exemplo, utilizando as preferências de outros usuários similares a este, assumindo de que suas preferências também sejam similares.

Estudo de áreas desprivilegiadas: Regiões que fornecem uma pequena quantidade de dados em comparação com outras regiões de uma mesma cidade pode indicar uma falta de acesso à tecnologia por parte da população, uma vez que o uso de RSPs geralmente necessita *smartphones* e planos de dados 3G ou 4G, que são caros em alguns países. Os resultados preliminares sobre o uso de RSPs nesses cenários demonstram boas oportunidades para a visualização de fatos interessantes. Por exemplo, analisando os dados para a cidade do Rio de Janeiro, mostrados na Figura 2.33, observa-se que é comum encontrar áreas muito pobres ao lado de áreas ricas. Repare que a atividade de sensoriamento é pequena nas áreas indicadas como áreas pobres. Esta informação pode ser útil de diversas formas, como para nortear melhores políticas públicas nessas áreas. A mesma informação pode ser obtida usando métodos tradicionais, tais como questionários, mas nessa possível nova maneira podemos ser capazes de obter esses dados de forma automática e mais barata usando RSPs. Para esse propósito, os algoritmos semelhantes ao proposto em [Cranshaw et al. 2012] poderiam ser aplicados.

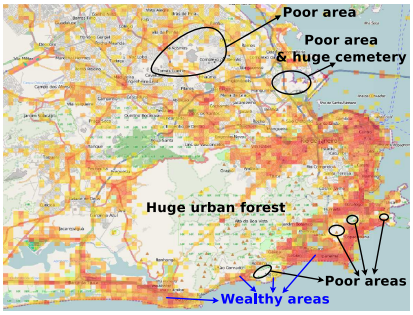


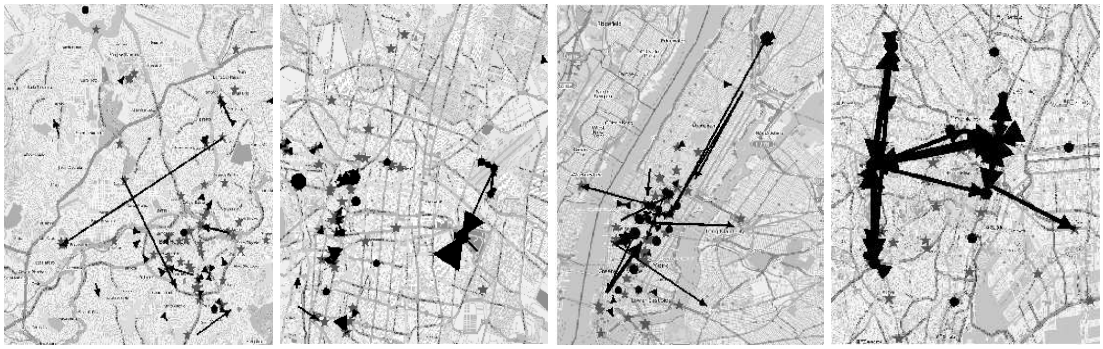
Figura 2.33. Exemplo de possível oportunidade para classificar áreas considerando a falta de dados [Silva et al. 2013a].

Classificação de áreas: Existem oportunidades de classificar áreas quando se considera conjuntamente o tempo e o local onde os *check-ins* são executadas. Pode ser possível visualizar multidões em uma cidade em tempo real. Além disso, os seres humanos possuem padrões sazonais, devido às suas rotinas. Esta sazonalidade tem um grande potencial para aplicações de previsão, uma vez que é muito provável que as pessoas repitam periodicamente suas atividades. Nós acreditamos que existem muitas oportunidades para previsão considerando o ritmo circadiano das pessoas, possibilitando a previsão, por exemplo, de multidões. Este tipo de informação é valiosa em muitos cenários, como em serviços para cidades inteligentes para evitar o tráfego em determinadas áreas e oferecer rotas alternativas para os usuários. Por exemplo, em [Hsieh et al. 2012] os autores propuseram um modelo que considera o tempo para recomendar rotas baseado em informações de RSPs para compartilhamento de localizações.

Além disso, com a utilização de dados sensoriados pelas pessoas através de RSPs é possível classificar áreas de diversas maneiras. Algumas delas foram discutidas aqui, como com relação ao cheiro, ruído e aspectos visuais. Isso pode ser útil para diversas novas aplicações e serviços. Um exemplo seria uma nova ferramenta de sugestão de rotas que sugere o menor caminho que é também o mais olfativamente agradável (as pessoas que praticam corrida de rua podem querer evitar ruas com fortes níveis de emissão de gases).

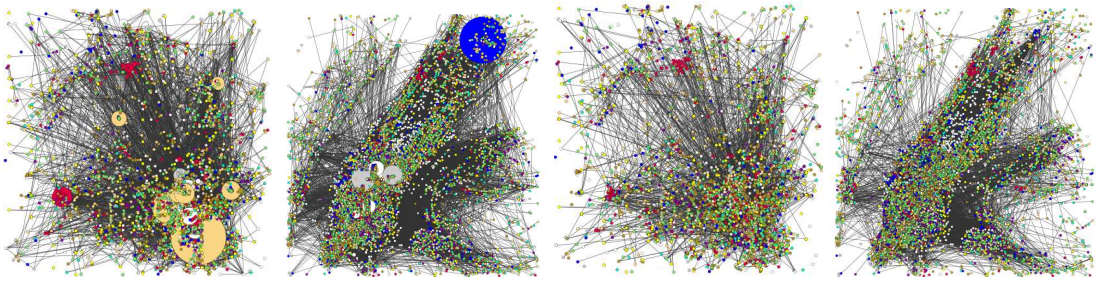
Grafos de transição urbana: Figura 2.34 mostra arestas ponderadas e nós impor-

tantes (50 maiores pesos de arestas e graus de nó) para Belo Horizonte, Cidade do México, Nova Iorque e Tóquio (considerando grafos de transição urbana, como os apresentados na seção 2.6.1, só que os nós agora são localidades individuais, ao invés de categorias de locais). Estrelas representam locais bastante visitados, setas pretas representam as arestas e círculos pretos representam *self-loops*. Quanto maior o símbolo, maior o valor. Repare que para Belo Horizonte e Cidade do México a maioria das arestas importantes são *self-loops* e de baixa distância, o que implica que as pessoas tendem a realizar atividades no bairro onde elas estão. Por outro lado, para Nova Iorque e Tóquio, cidades que são conhecidas por seus rápidos sistemas de transporte público, existe a tendência de algumas arestas ponderadas pesadas de longa distância ao longo de ligações de transporte público. Repare o potencial para serviços que dão suporte ao planejamento urbano.



(a) Belo Horizonte (b) Cidade do México (c) Nova Iorque (d) Tóquio

Figura 2.34. Arestas e nós importantes nos grafos de transição urbana de diversas cidades [Silva et al. 2013a].



(a) BH - Day (b) NY - Day (c) BH - Night (d) NY - Night

Figura 2.35. *Betweenness* dos nós de grafos de transição urbana para Nova Iorque e Belo Horizonte [Silva et al. 2014d].

Nessa mesma direção, podemos estudar métricas de centralidade nessa rede. Por exemplo, na Figura 2.35, mostramos os valores de centralidade de intermediação para os nós da rede. Cada cor é relacionada com uma categoria de local, e o tamanho do símbolo reflete a proporção do valor calculado. Esta abordagem pode ser utilizada para apoiar diversas aplicações, por exemplo, se for verificado um fluxo incomum e constante de pessoas entre dois locais de negócios independentes em uma cidade, os proprietários poderiam assinar um acordo comercial para aumentar suas receitas, como fazendo propaganda entre suas empresas.

Além disso, a técnica City Image, mencionada anteriormente, pode ser expandida para considerar subcategorias de locais, em vez de categorias principais. Como os dados de RSPs são altamente enviesados, algumas das transições mais populares entre as subcategorias devem ser bons indicadores da dinâmica da cidade. Essa técnica pode ser útil como uma forma de medir a distância entre duas cidades, permitindo a comparação de cidades e um agrupamento mundial que poderia ser interessante para sistemas de recomendação.

2.10. Conclusão

Neste minicurso, discutimos o conceito de computação urbana. Mostramos a relevância da área e motivamos a construção de novas aplicações para tratar questões relacionadas com a dinâmica da cidade e do comportamento social urbano. Discutimos também o sensoriamento urbano com redes de sensoriamento participativo. Mostramos que as RSPs oferecem oportunidades sem precedentes de acesso a dados de sensoriamento em escala planetária, dados que nos ajudam a entender melhor sociedades urbanas.

Estudamos também as principais técnicas utilizadas em trabalhos da área de computação urbana, bem como algumas das tecnologias e ferramentas comumente utilizadas para a análise de dados urbanos. Além disso, apresentamos diversos desafios sobre tópicos de pesquisa atuais relacionados com computação urbana utilizando dados de RSPs. Ressaltamos também várias oportunidades relacionadas ao uso de dados de RSPs em novos serviços e aplicações da área de computação urbana.

Agradecimentos

Gostaríamos de agradecer a fundamental ajuda de Pedro O. S. Vaz de Melo e Jussara M. Almeida para a realização deste trabalho. Gostaríamos de agradecer também aos alunos: Vinícius Mota, João Borges Neto, Clayson Celes, Felipe Cunha, Anna Ribeiro, Ana Ferreira e Virginia Kestering. Este trabalho é financiado parcialmente com o apoio das agências: CNPq, CAPES, FAPEMIG e Fundação Araucária.

Referências

[Adomavicius and Tuzhilin 2005] Adomavicius, G. and Tuzhilin, A. (2005). Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *Knowledge and Data Engineering, IEEE Transactions on*, 17(6):734–749.

[Alstott et al. 2014] Alstott, J., Bullmore, E., and Plenz, D. (2014). powerlaw: A python package for analysis of heavy-tailed distributions. *PLoS ONE*, 9(1):e85777.

[Balcan et al. 2009] Balcan, D., Colizza, V., Gonçalves, B., Hu, H., Ramasco, J. J., and Vespignani, A. (2009). Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences*, 106(51):21484–21489.

[Barbosa et al. 2014] Barbosa, L., Pham, K., Silva, C., Vieira, M. R., and Freire, J. (2014). Structured open urban data: understanding the landscape. *Big data*, 2(3):144–154.

[Barth 1969] Barth, F. (1969). *Ethnic groups and boundaries: the social organization of culture difference*. Scandinavian university books. Little, Brown.

[Benevenuto et al. 2011] Benevenuto, F., Almeida, J. M., and Silva, A. S. (2011). Explorando redes sociais online: Da coleta e análise de grandes bases de dados às aplicações. *Proc. of SBRC'11*, pages 63–94.

[Blei et al. 2003] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

[Bollen et al. 2011] Bollen, J., Mao, H., and Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8.

[Brenner and Smith 2013] Brenner, J. and Smith, A. (2013). 72% of online adults are social networking site users. http://goo.gl/HTgNy3.

[Brockmann et al. 2006] Brockmann, D., Hufnagel, L., and Geisel, T. (2006). The scaling laws of human travel. *Nature*, 439(7075):462–465.

[Brush et al. 2010] Brush, A. B., Krumm, J., and Scott, J. (2010). Exploring end user preferences for location obfuscation, location-based services, and the value of location. In *Proc. of Ubicomp '10*, pages 95–104, Copenhagen, Denmark. ACM.

[Burke et al. 2006] Burke, J., Estrin, D., Hansen, M., Parker, A., Ramanathan, N., Reddy, S., and Srivastava, M. B. (2006). Participatory sensing. In *Proc. of Workshop on World-Sensor-Web (WSW'06)*, pages 117–134, Boulder, USA.

[Burt 1992] Burt, R. S. (1992). *Structural Holes: The Social Structure of Competition*. Harvard University Press.

[CENS/UCLA] CENS/UCLA. *Participatory Sensing / Urban Sensing Projects*. http://research.cens.ucla.edu/.

[Cha et al. 2010] Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, K. (2010). Measuring user influence in twitter: The million follower fallacy. In *Proc. of ICWSM'10*, Washington, USA.

[Chatterjee and Hadi 2015] Chatterjee, S. and Hadi, A. S. (2015). *Regression analysis by example*. John Wiley & Sons.

[Cheng et al. 2011] Cheng, Z., Caverlee, J., Lee, K., and Sui, D. Z. (2011). Exploring Millions of Footprints in Location Sharing Services. In *Proc. of ICWSM'11*, Barcelona, Spain.

[Cho et al. 2011] Cho, E., Myers, S. A., and Leskovec, J. (2011). Friendship and mobility: user movement in location-based social networks. In *Proc. KDD '11*, pages 1082–1090, San Diego, USA. ACM.

[Clauset et al. 2009] Clauset, A., Shalizi, C. R., and Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Rev.*, 51(4):661–703.

[Coen-Porisini and Sicari 2012] Coen-Porisini, A. and Sicari, S. (2012). Improving data quality using a cross layer protocol in wireless sensor networks. *Comput. Netw.*, 56(17):3655–3665.

[Crandall et al. 2009] Crandall, D. J., Backstrom, L., Huttenlocher, D., and Kleinberg, J. (2009). Mapping the world's photos. In *Proc. of WWW '09*, pages 761–770, Madrid, Spain. ACM.

[Cranshaw et al. 2012] Cranshaw, J., Schwartz, R., Hong, J. I., and Sadeh, N. (2012). The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City. In *Proc. of ICWSM'12*, Dublin, Ireland.

[Dey and Abowd 2000] Dey, A. K. and Abowd, G. D. (2000). Towards a Better Understanding of Context and Context-Awareness. In *Proc. of CHI 2000 Workshops*, The Hague, The Netherlands.

[D'Hondt et al. 2013] D'Hondt, E., Stevens, M., and Jacobs, A. (2013). Participatory noise mapping works! an evaluation of participatory sensing as an alternative to standard techniques for environmental monitoring. *Pervasive and Mobile Computing*, 9(5):681–694.

[Duggan and Smith 2014] Duggan, M. and Smith, A. (2014). Social media update 2013. http://goo.gl/JhuiOG.

[Easley and Kleinberg 2010] Easley, D. and Kleinberg, J. (2010). *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge University Press.

[Eric Paulos and Townsend 2004] Eric Paulos, K. A. and Townsend, A. (2004). Ubicomp in the urban frontier. In *Workshop at Ubicomp'04*, Nottingham, UK.

[Ester et al. 1996] Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proc. of KDD-96*, Portland, USA.

[Freedman 2009] Freedman, D. A. (2009). *Statistical models: theory and practice*. cambridge university press.

[Ganti et al. 2011] Ganti, R., Ye, F., and Lei, H. (2011). Mobile crowdsensing: current state and future challenges. *Communications Magazine, IEEE*, 49(11):32 –39.

[Gao et al. 2015] Gao, H., Liu, C., Wang, W., Zhao, J., Song, Z., Su, X., Crowcroft, J., and Leung, K. (2015). A survey of incentive mechanisms for participatory sensing. *Communications Surveys Tutorials, IEEE*, PP(99):1–1.

[Gao et al. 2013] Gao, H., Tang, J., Hu, X., and Liu, H. (2013). Exploring temporal effects for location recommendation on location-based social networks. In *Proc. of RecSys '13*, pages 93–100, Hong Kong, China.

[Giannotti et al. 2012] Giannotti, F., Pedreschi, D., Pentland, A., Lukowicz, P., Kossmann, D., Crowley, J., and Helbing, D. (2012). A planetary nervous system for social mining and collective awareness. *The Eur. Phy. Jour. Special Topics*, 214(1):49–75.

[Gomide et al. 2011] Gomide, J., Veloso, A., Jr., W. M., Almeida, V., Benevenuto, F., Ferraz, F., and Teixeira, M. (2011). Dengue surveillance based on a computational model of spatio-temporal locality of twitter. In *Proc. of WebSci'11*, Evanston, USA.

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

[Gonçalves et al. 2013] Gonçalves, P., Araújo, M., Benevenuto, F., and Cha, M. (2013). Comparing and combining sentiment analysis methods. In *Proceedings of the 1st ACM Conference on Online Social Networks (COSN'13)*, Boston, USA.

[Harrell 2013] Harrell, F. E. (2013). *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. Springer Science & Business Media.

[Hartigan and Wong 1979] Hartigan, J. A. and Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Applied statistics*, pages 100–108.

[Hochman and Schwartz 2012] Hochman, N. and Schwartz, R. (2012). Visualizing instagram: Tracing cultural visual rhythms. In *Proc. of ICWSM'12*, pages 6–9, Dublin, Ireland. AAAI.

[Hsieh et al. 2012] Hsieh, H.-P., Li, C.-T., and Lin, S.-D. (2012). Exploiting large-scale check-in data to recommend time-sensitive routes. In *Proc. of UrbComp '12*, pages 55–62, Beijing, China. ACM.

[Instagram 2014] Instagram (2014). Instagram today: 200 million strong. http://blog.instagram.com/post/80721172292/200m.

[Jolliffe 2002] Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer, second edition.

[Joseph et al. 2012] Joseph, K., Tan, C. H., and Carley, K. M. (2012). Beyond local, categories and friends: clustering foursquare users with latent topics. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 919–926, Pittsburgh, USA. ACM.

[Jung 2015] Jung, M. (2015). *Essential Python Geospatial Libraries*. Indendent Group. `https://github.com/SpatialPython/spatial_python/blob/master/packages.md`.

[Kahle and Wickham 2013] Kahle, D. and Wickham, H. (2013). ggmap: Spatial visualization with ggplot2. *The R Journal*, 5(1):144–161.

[Karamshuk et al. 2013] Karamshuk, D., Noulas, A., Scellato, S., Nicosia, V., and Mascolo, C. (2013). Geo-spotting: Mining online location-based services for optimal retail store placement. In *Proc. of KDD '13*, pages 793–801, Chicago, Illinois, USA. ACM.

[Kaufman and Rousseeuw 2009] Kaufman, L. and Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons.

[Kindberg et al. 2007] Kindberg, T., Chalmers, M., and Paulos, E. (2007). Guest editors' introduction: Urban computing. *IEEE Pervasive Computing*, 6(3):18–20.

[Kisilevich et al. 2010] Kisilevich, S., Krstajic, M., Keim, D., Andrienko, N., and Andrienko, G. (2010). Event-based analysis of people's activities and behavior using flickr and panoramio geotagged photo collections. In *Proc. of Conf. on Inf. Vis.*, pages 289–296, London, UK. IEEE.

[Kostakos 2009] Kostakos, V. (2009). Temporal graphs. *Physica A: Statistical Mechanics and its Applications*, 388(6):1007–1023.

[Kostakos and O'Neill 2008] Kostakos, V. and O'Neill, E. (2008). {City ware: Urban Computing to Bridge Online}. *Handbook of Research on Urban Informatics: The Practice and Promise of the Real-Time City*.

[Krings et al. 2009] Krings, G., Calabrese, F., Ratti, C., and Blondel, V. D. (2009). Urban gravity: a model for inter-city telecommunication flows. *Journal of Statistical Mechanics: Theory and Experiment*, 2009(07):L07003.

[Lane et al. 2010] Lane, N., Miluzzo, E., Lu, H., Peebles, D., Choudhury, T., and Campbell, A. (2010). A survey of mobile phone sensing. *Comm. Mag., IEEE*, 48(9):140 –150.

[Lane et al. 2008] Lane, N. D., Eisenman, S. B., Musolesi, M., Miluzzo, E., and Campbell, A. T. (2008). Urban sensing systems: Opportunistic or participatory? In *Proc. of HotMobile '08*, pages 11–16, Napa Valley, California. ACM.

[Lee and Hoh 2010] Lee, J.-S. and Hoh, B. (2010). Dynamic pricing incentive for participatory sensing. *Pervasive and Mobile Computing*, 6(6):693–708.

[Loureiro et al. 2003] Loureiro, A. A. F., Nogueira, J. M. S., Ruiz, L. B., Mini, R. A., Nakamura, E. F., and Figueiredo, C. M. S. (2003). Redes de sensores sem fio. *Proc. of SBRC'03*, pages 179–226.

[Luxburg 2007] Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416.

[MacKerron and Mourato 2013] MacKerron, G. and Mourato, S. (2013). Happiness is greater in natural environments. *Global Environmental Change*, 23(5):992–1000.

[Maisonneuve et al. 2009] Maisonneuve, N., Stevens, M., Niessen, M. E., and Steels, L. (2009). Noisetube: Measuring and mapping noise pollution with mobile phones. In *Information Technologies in Environmental Engineering*, pages 215–228. Springer.

[Martine et al. 2007] Martine, G., Marshall, A., et al. (2007). State of world population 2007: unleashing the potential of urban growth. In *State of world population 2007: unleashing the potential of urban growth*. UNFPA.

[Mashhadi and Capra 2011] Mashhadi, A. J. and Capra, L. (2011). Quality Control for Real-time Ubiquitous Crowdsourcing. In *Proc. of UbiCrowd'11*, pages 5–8, Beijing, China.

[Nazir et al. 2008] Nazir, A., Raza, S., and Chuah, C.-N. (2008). Unveiling facebook: A measurement study of social network based applications. In *Proc. of IMC '08*, pages 43–56, Vouliagmeni, Greece.

[Newman 2010] Newman, M. (2010). *Networks: an introduction*. Oxford University Press, Inc.

[Newman 2003] Newman, M. E. J. (2003). The Structure and Function of Complex Networks. *SIAM Review*, 45(2):167–256.

[Ng et al. 2002] Ng, A. Y., Jordan, M. I., Weiss, Y., et al. (2002). On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 2:849–856.

[Nguyen and Szymanski 2012] Nguyen, T. and Szymanski, B. K. (2012). Using location-based social networks to validate human mobility and relationships models. *arXiv preprint arXiv:1208.3653*.

[Nicosia et al. 2013] Nicosia, V., Tang, J., Mascolo, C., Musolesi, M., Russo, G., and Latora, V. (2013). Graph metrics for temporal networks. In *Temporal Networks*, pages 15–40. Springer.

[Noulas et al. 2011a] Noulas, A., Scellato, S., Mascolo, C., and Pontil, M. (2011a). An Empirical Study of Geographic User Activity Patterns in Foursquare. In *Proc. of ICWSM'11*, Barcelona, Spain.

[Noulas et al. 2011b] Noulas, A., Scellato, S., Mascolo, C., and Pontil, M. (2011b). Exploiting Semantic Annotations for Clustering Geographic Areas and Users in Location-based Social Networks. In *Proc. of the Fifth Int'l Conf. on Weblogs and Social Media (ICWSM'11)*, Barcelona, Spain. AAAI.

[Paulos and Goodman 2004] Paulos, E. and Goodman, E. (2004). The familiar stranger: anxiety, comfort, and play in public places. In *Proc. of CHI'04*, pages 223–230, Vienna, Austria. ACM.

[Poblete et al. 2011] Poblete, B., Garcia, R., Mendoza, M., and Jaimes, A. (2011). Do all birds tweet the same?: characterizing twitter around the world. In *Proc. of CIKM*, pages 1025–1030, Glasgow, UK. ACM.

[Pokorny 2013] Pokorny, J. (2013). Nosql databases: a step to database scalability in web environment. *International Journal of Web Information Systems*, 9(1):69–82.

[Pontes et al. 2012] Pontes, T., Magno, G., Vasconcelos, M., Gupta, A., Almeida, J., Kumaraguru, P., and Almeida, V. (2012). Beware of what you share: Inferring home location in social networks. In *Proc. of ICDMW*, pages 571–578, Brussels, Belgium.

[Quercia et al. 2012] Quercia, D., Capra, L., and Crowcroft, J. (2012). The social world of twitter: Topics, geography, and emotions. In *Proc. of ICWSM'12*, Dublin, Ireland.

[Quercia et al. 2014] Quercia, D., Schifanella, R., and Aiello, L. M. (2014). The shortest path to happiness: Recommending beautiful, quiet, and happy routes in the city. In *Proceedings of the 25th ACM Conference on Hypertext and Social Media*, HT '14, pages 116–125, New York, NY, USA. ACM.

[Quercia et al. 2015] Quercia, D., Schifanella, R., Aiello, L. M., and McLean, K. (2015). Smelly maps: The digital life of urban smellscapes. In *Proc. of ICWSM'15*, Oxford, UK.

[Reddy et al. 2010] Reddy, S., Estrin, D., Hansen, M., and Srivastava, M. (2010). Examining micro-payments for participatory sensing data collections. In *Proc. of Ubicomp '10*, pages 33–36, Copenhagen, Denmark. ACM.

[Reis et al. 2015] Reis, J., Benevenuto, F., Vaz de Melo, P., Prates, R., Kwak, H., and An, J. (2015). Breaking the news: First impressions matter on online news. In *Proceedings of the 9th International AAAI Conference on Web-Blogs and Social Media*, Oxford, UK.

[Sakaki et al. 2010a] Sakaki, T., Okazaki, M., and Matsuo, Y. (2010a). Earthquake shakes twitter users: real-time event detection by social sensors. In *Proc. of WWW'10*, pages 851–860, Raleigh, USA.

[Sakaki et al. 2010b] Sakaki, T., Okazaki, M., and Matsuo, Y. (2010b). Earthquake shakes twitter users: real-time event detection by social sensors. In *Proc. of WWW'10*, pages 851–860, Raleigh, USA. IW3C2.

[Shannon 1948] Shannon, C. E. (1948). A mathematical theory of communication. *Bell system tech. jour.*, 27.

[Silva et al. 2014a] Silva, T., Vaz De Melo, P., Almeida, J., and Loureiro, A. (2014a). Large-scale study of city dynamics and urban social behavior using participatory sensing. *Wireless Communications, IEEE*, 21(1):42–51.

[Silva et al. 2014b] Silva, T. H., Vaz de Melo, P., Almeida, J., Viana, A., Salles, J., and Loureiro, A. (2014b). Participatory Sensor Networks as Sensing Layers. In *Proc. of SocialCom'14*, Sydney, Australia.

[Silva et al. 2012] Silva, T. H., Vaz de Melo, P. O. S., Almeida, J. M., and Loureiro, A. A. F. (2012). Visualizing the invisible image of cities. In *Proc. IEEE CPScom'12*, pages 382–389, Besancon, France.

[Silva et al. 2013a] Silva, T. H., Vaz de Melo, P. O. S., Almeida, J. M., and Loureiro, A. A. F. (2013a). Challenges and opportunities on the large scale study of city dynamics using participatory sensing. In *Proc. of IEEE ISCC'13*, pages 528–534, Split, Croatia.

[Silva et al. 2013b] Silva, T. H., Vaz de Melo, P. O. S., Almeida, J. M., and Loureiro, A. A. F. (2013b). Uma Fotografia do Instagram: Caracterização e Aplicação. In *Proc. of SBRC'13*, Brasília, Brazil.

[Silva et al. 2014c] Silva, T. H., Vaz de Melo, P. O. S., Almeida, J. M., Musolesi, M., and Loureiro, A. A. F. (2014c). You are What you Eat (and Drink): Identifying Cultural Boundaries by Analyzing Food & Drink Habits in Foursquare. In *Proc. of ICWSM'14*, Ann Arbor, USA.

[Silva et al. 2013c] Silva, T. H., Vaz de Melo, P. O. S., Almeida, J. M., Salles, J., and Loureiro, A. A. F. (2013c). A picture of Instagram is worth more than a thousand words: Workload characterization and application. In *Proc. of DCOSS'13*, pages 123–132, Cambridge, USA.

[Silva et al. 2013d] Silva, T. H., Vaz de Melo, P. O. S., Almeida, J. M., Salles, J., and Loureiro, A. A. F. (2013d). A comparison of foursquare and instagram to the study of city dynamics and urban social behavior. In *Proc. of UrbComp'13*, pages 1–8, Chicago, USA.

[Silva et al. 2014d] Silva, T. H., Vaz de Melo, P. O. S., Almeida, J. M., Salles, J., and Loureiro, A. A. F. (2014d). Revealing the city that we cannot see. *ACM Trans. Internet Technol.*, 14(4):26:1–26:23.

[Silva et al. 2013e] Silva, T. H., Vaz de Melo, P. O. S., Viana, A., Almeida, J. M., Salles, J., and Loureiro, A. A. F. (2013e). Traffic Condition is more than Colored Lines on a Map: Characterization of Waze Alerts. In *Proc. of SocInfo'13*, pages 309–318, Kyoto, Japan.

[Sinnott 1984] Sinnott, R. W. (1984). Virtues of the Haversine. *Sky and Telescope*, 68(2):159+.

[Sørensen 1948] Sørensen, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. *Biologiske Skrifter*, 5(4).

[Srivastava et al. 2012] Srivastava, M., Abdelzaher, T., and Szymanski, B. (2012). Human-centric sensing. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 370(1958):176–197.

[Sutton et al. 2009] Sutton, T., Dassau, O., and Sutton, M. (2009). *A gentle introduction to gis*.

[Thelwall et al. 2010] Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., and Kappas, A. (2010). Sentiment in short strength detection informal text. *J. Am. Soc. Inf. Sci. Technol.*, 61(12):2544–2558.

[Toch et al. 2010] Toch, E., Cranshaw, J., Drielsma, P. H., Tsai, J. Y., Kelley, P. G., Springfield, J., Cranor, L., Hong, J., and Sadeh, N. (2010). Empirical models of privacy in location sharing. In *Proc. of Ubicomp'10*, pages 129–138, Copenhagen, Denmark. ACM.

[Tostes et al. 2013] Tostes, A. I. J., Duarte-FIgueiredo, F., Assunção, R., Salles, J., and Loureiro, A. A. F. (2013). From data to knowledge: City-wide traffic flows analysis and prediction using bing maps. In *Proc. of ACM UrbComp'13*, Chicago, USA.

[Tostes et al. 2014] Tostes, A. I. J., Silva, T. H., Duarte-FIgueiredo, F., and Loureiro, A. A. F. (2014). Studying traffic conditions by analyzing foursquare and instagram data. In *Proc. of ACM PE-WASUN'14*, Montreal, Canada.

[Yan 2009] Yan, X. (2009). *Linear regression analysis: theory and computing*. World Scientific.

[Youyou et al. 2015] Youyou, W., Kosinski, M., and Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Nat. Acad. of Sci.*, 112(4):1036–1040.

[Yuan et al. 2013] Yuan, Q., Cong, G., Ma, Z., Sun, A., and Thalmann, N. M. (2013). Time-aware point-of-interest recommendation. In *Proc. of SIGIR '13*, pages 363–372, Dublin, Ireland. ACM.

[Zaki and Meira Jr 2014] Zaki, M. J. and Meira Jr, W. (2014). *Data mining and analysis: fundamental concepts and algorithms*. Cambridge University Press.

[Zhang et al. 2013] Zhang, K., Jin, Q., Pelechrinis, K., and Lappas, T. (2013). On the importance of temporal dynamics in modeling urban activity. In *Proc. of UrbComp'13*, pages 7:1–7:8, Chicago, Illinois.

[Zheng et al. 2014a] Zheng, Y., Capra, L., Wolfson, O., and Yang, H. (2014a). Urban computing: concepts, methodologies, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(3):38.

[Zheng et al. 2014b] Zheng, Y., Liu, T., Wang, Y., Zhu, Y., Liu, Y., and Chang, E. (2014b). Diagnosing new york city's noises with ubiquitous data. In *Proc. of UbiComp'14*, pages 715–725, Seattle, Washington. ACM.

[Zheng et al. 2009] Zheng, Y., Zhang, L., Xie, X., and Ma, W.-Y. (2009). Mining interesting locations and travel sequences from gps trajectories. In *Proc. of WWW'09*, pages 791–800, Madrid, Spain. ACM.

[Zheng et al. 2012] Zheng, Y.-T., Zha, Z.-J., and Chua, T.-S. (2012). Mining travel patterns from geotagged photos. *ACM Trans. Intell. Syst. Technol.*, 3(3):56:1–56:18.

[Zipf 1946] Zipf, G. K. (1946). The p1 p2/d hypothesis: On the intercity movement of persons. *American sociological review*, pages 677–686.