UNIVERSIDADE ESTADUAL PAULISTA

"Júlio de Mesquita Filho"

Pós-Graduação em Ciência da Computação

Rafael Mariano Christófano

PlaceProfile: Empregando análise visual e técnicas de agrupamento para criar perfis de regiões com base em pontos de interesse.

Presidente Prudente
2021

Rafael Mariano Christófano

PlaceProfile: Empregando análise visual e técnicas de agrupamento para criar perfis de regiões com base em pontos de interesse

Orientador: Prof. Dr. Danilo Medeiros Eler

Dissertação apresentada como parte dos requisitos para obtenção do título de Mestre em Ciência da Computação – Área de Concentração em Computação Aplicada, junto ao Programa de Pós-Graduação em Ciência da Computação, da Faculdade de Ciências e Tecnologia da Universidade Estadual Paulista "Júlio de Mesquita Filho", Câmpus de Presidente Prudente.

Presidente Prudente
2021

C556p

Christófano, Rafael

PlaceProfile: Empregando análise visual e técnicas de agrupamento para criar perfis de regiões com base em pontos de interesse. / Rafael Christófano. -- Presidente Prudente, 2021

83 p.

Dissertação (mestrado) - Universidade Estadual Paulista (Unesp), Faculdade de Ciências e Tecnologia, Presidente Prudente

Orientador: Danilo Eler

1. Visualização de Informação. 2. Pontos de Interesse. 3. Cidades Inteligentes. 4. Mobilidade Inteligente. 5. Algoritmo de Agrupamento. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca da Faculdade de Ciências e Tecnologia, Presidente Prudente. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

UNIVERSIDADE ESTADUAL PAULISTA

Câmpus de Presidente Prudente

CERTIFICADO DE APROVAÇÃO

TÍTULO DA DISSERTAÇÃO: PlaceProfile: Empregando análise visual e técnicas de agrupamento para criar perfis de regiões com base em pontos de interesse,

AUTOR: RAFAEL MARIANO CHRISTOFANO

ORIENTADOR: DANILO MEDEIROS ELER

Aprovado como parte das exigências para obtenção do Título de Mestre em CIÊNCIA DA COMPUTAÇÃO, área: Computação Aplicada pela Comissão Examinadora:

Prof. Dr. DANILO MEDEIROS ELER (Participaçao Virtual)
Departamento de Matemática e Computação / UNESP/Câmpus de Presidente Prudente

Prof. Dr. DANILLO ROBERTO PEREIRA (Participaçao Virtual)
Faculdade de Informática de Presidente Prudente / Universidade do Oeste Paulista - UNOESTE

Prof. Dr. JOSE REMO FERREIRA BREGA (Participaçao Virtual)
Departamento de Computação / UNESP/Câmpus de Bauru

Presidente Prudente, 05 de março de 2021

A minha esposa Juliana e a meus filhos Giovanna e Gabriel.
E aos meus pais.

AGRADECIMENTOS

Primeiramente, agradeço a Deus por me guiar e abrir as portas certas nos momentos certos durante esses anos de estudo. Agradeço a minha esposa Juliana e a meus filhos Giovanna e Gabriel, que muitas vezes tiveram que entender uma frase que disse repetidas vezes que "agora o papai não pode". Aos meus pais que são o alicerce de tudo que eu conquistei nessa vida. E, por fim, ao meu professor e orientador Danilo Eler que acreditou em mim desde o início, mesmo sabendo da dura jornada pela frente.

"Muitas vezes esperamos por um grande milagre que mudará nossas vidas, mas esquecemos tais milagres acontecem diariamente, basta observar." – Próprio autor.

RESUMO

Entender como as atividades comerciais e sociais, bem como os pontos de interesse estão localizados em uma cidade, é essencial para planejar cidades eficientes em termos de mobilidade inteligente. Ao longo dos anos, o crescimento das fontes de dados de distintas redes sociais online têm permitido novas perspectivas para aplicativos que fornecem mecanismos para ajudar na compreensão de como as pessoas se deslocam entre as diferentes regiões de uma cidade. Para apoiar empresas e governos para melhor compreender e comparar regiões distintas de uma cidade, este trabalho propõe uma aplicação web chamado PlaceProfile para criação de perfis visuais de áreas de uma cidade com base em uma visualização iconográfica e para rotular áreas baseadas em algoritmos de agrupamento. Os resultados da visualização são sobrepostos no Google Maps para enriquecer o layout do mapa e a análise ajuda a entender o perfil da região em um relance. Além disso, o PlaceProfile coordena um gráfico de radar com áreas selecionadas pelo usuário para permitir a inspeção detalhada da frequência das categorias de pontos de interesse (POIs). Esta abordagem de visualizações coordenadas também apoia a explicabilidade de algoritmos de agrupamento por fornecer inspeções dos atributos usados para calcular semelhanças, ou seja, o número de pontos de interesse em cada categoria. Assim, uma variedade de pesquisas e aplicações voltadas a resolver problemas de mobilidade urbana podem se beneficiar dos resultados produzidos pelo PlaceProfile.

Palavras-chave: Perfil de áreas, cidades inteligentes, mobilidade inteligente, algoritmos de agrupamento, POIs, visualização.

ABSTRACT

Understanding how commercial and social activities as well as points of interest are located in a city is essential to plan efficient cities in terms of smart mobility. Over the years, the growth of data sources from distinct online social networks have enabled new perspectives to applications that provide mechanisms to aid in comprehension of how people displace between different regions within a city. To support enterprises and governments to better understand and compare distinct regions of a city, this work proposes a web application called PlaceProfile to perform visual profiling of city areas based on a iconographic visualization and to label areas based on clustering algorithms. The visualization results are overlayered on Google Maps to enrich the map layout and aid analyst to understand region profiling at a glance. In addition, PlaceProfile coordinates a radar chart with areas selected by the user to enable detailed inspection of the frequency of categories of points of interest (POIs). This linked views approach also supports explainability of clustering algorithms by providing inspections of the attributes used to compute similarities, that is, the number of points of interest in each category. A variety of research and applications aimed at urban mobility problems can benefit from the results obtained by PlaceProfile.

Keywords: Profilling areas, smart cities, smart mobility, clustering algorithm, POIs, visualization.

LISTA DE FIGURAS

Figura 1 – Processo de extração de conhecimento inicia na obtenção de dados brutos coletados de diversos meios, em seguida é realizado o pre-processamento em que os dados são limpos e organizados para que possa ser utilizados apenas os dados de interesse. Os dados interessantes são enfim minerados para que o processo finalize e permita ao usuário obter informações relevantes. 6
Figura 2 – Uma visão Geral dos métodos analisados no trabalho de Marjani et al. (2017). 9
Figura 3 – Elementos no processo da interpretação da visão humana (O'CONNOR, 2015). 16
Figura 4 – Arranjo básico de subjanelas de dados com seis dimensões, adaptado para o português do trabalho de Keim (2000). 18
Figura 5 – Exemplos de técnicas de visualização iconográfica. (a) ícones de agulha (WARD, 1994), (b) rostos (ASTEL et al., 2006), (c) ícones de bonecos de palito (KAISER, 2000), (d) ícones de estrela (JMP, 2021), (e) ícones coloridos (NOCKE; SCHLECHTWEG; SCHUMANN, 2005), (f) Mapas de grades (DATAVIZCATALOGUE, 2021). 19
Figura 6 – Visualização da técnica de hierarquia dimensional de dados (KEIM, 2002). 20
Figura 7 – Visualização da técnica de Coordenadas Paralelas (KEIM, 2002). . . . 21
Figura 8 – Um dos primeiros trabalhos sobre monitoramento do tráfego urbano, na Figura (a) os fluxos de dados são analisados espacialmente, em (b) tematicamente com gráficos de áreas, em (c) temporalmente, através de gráficos de linhas que representam séries temporais e (d) com agregação, usando gráfico de barras para mostrar intervalo entre valores (CLARAMUNT; JIANG; BARGIELA, 2000). 22
Figura 9 – Uma amostra dos POIs, fornecidos pelo *Google Maps*, na região central de Milão e a esquerda da imagem, mais informações sobre um ponto específico, a catedral de Milão (D'Andrea et al., 2018). 24
Figura 10 – O *pipeline* do processo da geração de valor através do *framework* proposto com o objetivo de identificar o perfil das diferentes regiões da cidade de Milão (D'Andrea et al., 2018). 24
Figura 11 – Resultado do agrupamento em um *heatmap* com 5 *clusters*, sobre o mapa da cidade de Milão (Fonte: *Google Maps*) com 1188 células com tamanho de 500 metros cada lado (D'Andrea et al., 2018). 26

Figura 12 – Um gráfico de radar com 4 *clusters* em escala logarítmica para a macro-categoria *DomCat*(D'Andrea et al., 2018). 27
Figura 13 – Pontos de registro da posição de um usuário em cada observação dos dados de CDR. É possível identificar 2 grupos de registros, isso demonstra que em cada evento de telefonia móvel em um mesmo local, foram geradas posições geográficas não tão precisas (WANG et al., 2010). . . 28
Figura 14 – Ilustrando o efeito da agregação de torres em super torres que cobrem áreas geográficas maiores. As coordenadas geográficas das torres foram agrupadas para produzir centróides espaciais que cobrem áreas maiores. Cada célula de voronoi é colorida com a atividade mais popular, explorando a popularidade de locais do *Foursquare* próximos (Noulas; Mascolo; Frias-Martinez, 2013). 29
Figura 15 – Padrão proposto por Bowman e Ben-Akiva (2001) relacionando as 23 atividades auto referidas em 9 grupos de atividades aplicadas no trabalho de (JIANG; FERREIRA; GONZÁLEZ, 2012). 30
Figura 16 – *Snapshots* obtidos da ferramenta de animação temporal demonstrando atividades humanas em diferentes horas do dia em um dia da semana em Chicago (JIANG; FERREIRA; GONZÁLEZ, 2012). 31
Figura 17 – *Pipeline* do processo de descoberta de conhecimento combinando dados de CDR com censitários de Cingapura (Jiang; Ferreira; Gonzalez, 2017). 32
Figura 18 – (a) Distribuição espacial dos fatores de expansão do usuário no nível da torre e distribuição de frequência (b) fatores de expansão do usuário e (c) fatores de expansão no dia do usuário. (Jiang; Ferreira; Gonzalez, 2017) 33
Figura 19 – Diagrama de rosas aplicado a entender o congestionamento do tráfego na cidade de Helsinque. Cada diagrama de rosas foi definido por um usuário especialista no trânsito de Helsinque em ponto estratégicos da cidade (ANDRIENKO et al., 2013). 34
Figura 20 – Visualização da matriz do mapa de calor para análise de congestionamento de tráfego (SONG; MILLER, 2012). 35
Figura 21 – Outras formas de representação do mapa de calor sobreposto em um mapa geográfico (POCO et al., 2015). 36
Figura 22 – Visualização apresenta a correlação de atividades urbanas entre cidades italianas em diferentes dias da semana (SAGL; LOIDL; BEINAT, 2012). 37
Figura 23 – Visualização baseada em mapas de fluxo de transferência de torres de telefonia móvel durante uma chamada telefônica ativa (DEMISSIE; CORREIA; BENTO, 2013). 38
Figura 24 – Visualização baseada em mapa de volume de entrega de entrada e saída usando círculos dimensionados (DEMISSIE; CORREIA; BENTO, 2013). 38

Figura 25 – Um sistema de análise visual proposto por Chen et al. (2015a) para análise visual de dados esparsos de *microblogging*. Várias técnicas de visualização estão inter-relacionadas. Gráficos de tempo (a) e matrizes de mapa de calor (c) representam a distribuição do movimento na distância e no tempo. A visualização baseada em mapas em (b) exibe fluxos de espaço e tempo agregados entre cidades. Finalmente, (d) apresenta o diagramas de Sankey para representar movimento pareados no tempo. . 39
Figura 26 – Ferramenta ICE mostrando os histogramas interativos e a interligação entre os mapas (PACK et al., 2009). 40
Figura 27 – PlaceProfile consiste em uma aplicação web para mineração visual de dados após a coleta dos pontos de interesse no Google Maps. Os principais componentes do PlaceProfile são preparação, coleta de dados, mineração de dados e a visualização. 41
Figura 28 – PlaceProfile: Neste exemplo, uma grid foi plotada sobre o mapa de uma área da cidade de São Paulo, cada célula da grid tem um tamanho de 250x250 metros, totalizando 476 células. 42
Figura 29 – PlaceProfile: Os dados coletados (pontos em vermelho) são plotados na grid em sua respectiva célula sobreposto ao mapa. 43
Figura 30 – PlaceProfile: Zoom aplicado sobre uma área de coleta com 18 células selecionadas. 44
Figura 31 – PlaceProfile: Ilustração da técnica de visualização aplicada para destacar as mais comuns macrocategorias por célula. Uma grid é sobreposta a uma região e cada célula da grid é dividida novamente em uma outra subgrid 10x10 (a) e as macrocategorias mais presentes nessa célula são ordenadas de cima para baixo de acordo com a quantidade de atividades relacionadas a macrocategoria (b). A última cor(marrom), está relacionada a soma de todas as outras macrocategorias identificadas na célula. 47
Figura 32 – PlaceProfile: Usando cores para codificar a macrocategoria mais comum em uma célula. Esta região é particularmente representada por macrocategorias de serviços (*services*) e lojas (*stores*). 47
Figura 33 – PlaceProfile: Análise de cluster. Com base nos recursos recuperados durante a etapa de coleta de dados, as células são agrupadas para ajudar na análise com base na similaridade. 48
Figura 34 – PlaceProfile: O gráfico Radar Chart mostra a proporção de macrocategorias para os dois clusters apresentados na Figura 33. Embora muito semelhantes em serviços (*services*), lojas (*stores*) e alimentação (*food*), esses dois grupos diferem nas macrocategorias cultural (*cultural*), de saúde *health* e financeira (*finance*). 49

Figura 35 – PlaceProfile: A grid plotada sobre o mapa de uma área da cidade de São Paulo, cada célula da grid tem um tamanho de 250x250 metros, totalizando 476 células. 50
Figura 36 – PlaceProfile: Os dados coletados (pontos em vermelho) são plotados na grid em sua respectiva célula e posição geográfica sobreposto ao mapa. 51
Figura 37 – PlaceProfile: Uma visão geral sobre a região central da cidade de São Paulo. Podemos ver principalmente uma divisão entre POIs relacionados a lojas (*stores*) e POIs relacionados a saúde (*health*). As células pretas correspondem a regiões com número de POIs abaixo do limite mínimo definido pelo usuário. 51
Figura 38 – PlaceProfile: Um comparativo do resultado visual demonstrando os resultados da análise iconográfica e para cada um dos algoritmos de agrupamento. Nesse cenário, é possível perceber que o Agglomerative destoou dos demais generalizando em um mesmo cluster grupos que os outros algoritmos classificaram em grupos separados. 53
Figura 39 – PlaceProfile: Análise de similaridade baseada em resultados de agrupamento usando o algoritmo K-means. O Radar Chart apoia na visualização mostrando que o cluster azul representa os POIs relacionados à macrocategoria loja (*store*), enquanto o cluster laranjas representa os POIs relacionados às macrocategorias de saúde (*health*), já as células em vermelho não há uma macrocategoria com grande destaque sobre as outras. 54
Figura 40 – PlaceProfile: Análise de similaridade baseada em resultados de agrupamento usando o algoritmo Agglomerative. O Radar Chart apoia na visualização mostrando que o cluster vermelho representa os POIs relacionados à macrocategoria loja (*store*), enquanto o cluster verde representa os POIs relacionados às macrocategorias de saúde (*health*), já as células em azul e laranja não há uma macrocategoria com grande destaque sobre as outras. 55
Figura 41 – PlaceProfile: Análise de similaridade baseada em resultados de agrupamento usando o algoritmo Fuzzy C-means. O Radar Chart apoia na visualização mostrando que os clusters na cor vermelho e azul representam os POIs relacionados à macrocategoria loja (*store*), porém o cluster azul também há uma incidência de atividades relacionados à macrocategoria serviços (*services*) e saúde (*health*), enquanto o cluster amarelo representa os POIs relacionados às macrocategorias de saúde (*health*), serviços (*services*) e lojas (*stores*), já as células em verde não há uma macrocategoria com grande destaque sobre as outras. 56

Figura 42 – PlaceProfile: Análise de similaridade baseada em resultados de agrupamento. As células selecionadas mostram que o cluster azul representa os POIs relacionados à macrocategoria loja (*store*), enquanto o cluster vermelho há um pequeno destaque para serviços (*service*), mas também há incidências de outras macrocategorias como de saúde (*health*), lojas (*stores*). 57
Figura 43 – PlaceProfile: Seleção de diferentes clusters para entender os padrões de POIs. O cluster verde parece ter POIs relacionados às macrocategorias de lojas (*stores*) e serviços (*services*), enquanto o cluster amarelo corresponde aos POIs relacionados à macrocategoria de saúde (*health*). 58

LISTA DE TABELAS

Tabela 1 – Matriz que relaciona os métodos de mineração de dados para aplicações em diversas áreas, o 'x' indica que a técnica de mineração pode ser usar para as respectivas aplicação, enquanto o '-' não é óbvio que a técnica pode ser usada. (Marjani et al., 2017). 10
Tabela 2 – Para melhor interpretação da representatividade de cada *cluster*, 5 macro-categorias foram criadas e suas medidas servem para identificar o perfil de cada região de Milão (D'Andrea et al., 2018). 26
Tabela 3 – Dados brutos coletados do Google Place 43
Tabela 4 – Tabela relacionando as diversas categorias de atividades similares capturadas do Google Places API, com suas respectivas macrocategorias, essa técnica foi usado no trabalho de D'Andrea et al. (2018). 45
Tabela 5 – Uma amostra da sumarização das macrocategorias por grupos(clusters): a contagem do total das 11 macrocategorias em cada célula cria os atributos que serão passados como parâmetro para o algoritmo de agrupamento e para a análise iconográfica 46

SUMÁRIO

Resumo . . . . . **vii**

Abstract . . . . . **viii**

Lista de Figuras . . . . . **ix**

Lista de Tabelas . . . . . **xiv**

1 INTRODUÇÃO **1**
1.1 Contextualização **2**
1.2 *Objetivos* **3**
1.3 *Organização* **4**

2 FUNDAMENTAÇÃO **6**
2.1 Extração de Conhecimento **6**
2.1.1 Obtenção de Dados Brutos 6
2.1.2 Mineração dos dados 8
2.2 Algoritmos de Agrupamento **11**
2.2.1 *K-Means* 11
2.2.2 *Fuzzy C-Means* 13
2.2.3 *Agglomerative* 14
2.3 Visualização da Informação **15**
2.3.1 Técnicas de Visualização 17
2.3.2 Técnicas Orientadas a Pixel 17
2.3.3 Técnicas Iconográficas 19
2.3.4 Técnicas Hierárquicas 20
2.3.5 Projeções Geométricas 20

3 TRABALHOS RELACIONADOS **22**
3.1 Identificar o Perfil de Regiões em uma Cidade **23**
3.2 Análise de Padrões de Atividade Humanas **30**
3.3 *Identificação de fluxos do tráfego de veículos* **34**
3.4 *Análise da dinâmica do deslocamento de pessoas* **36**
3.5 *Análise de Incidentes de tráfego* **37**

4 PLACEPROFILE: DESCOBRINDO PADRÕES BASEADOS EM PONTOS DE INTERESSE **41**

4.1 Preparação . . . 42
4.2 Coleta dos dados . . . 43
4.3 Mineração dos Dados . . . 44
4.4 Visualização . . . 46
4.5 Implementação . . . 49

5 RESULTADOS . . . 50
5.1 Análise Visual dos Perfis . . . 50
5.2 Análise Visual dos Algoritmos de Agrupamento . . . 52
5.2.1 Análise a partir de células selecionadas . . . 54

6 CONCLUSÃO E TRABALHOS FUTUROS . . . 59

REFERÊNCIAS . . . 61

1 INTRODUÇÃO

Com o crescimento das grandes cidades e o aumento da população mundial, o deslocamento das pessoas para realizar suas atividades diárias tem se tornado um grande desafio. Assim, criar soluções que melhorem o trânsito fazendo com que as pessoas possam se deslocar de um ponto a outro de uma forma ágil e segura tem sido um desafio para os governantes locais administrarem (D'Andrea et al., 2018). Portanto, o planejamento das cidades está intimamente ligado à mobilidade humana em um território urbano, refletindo diretamente no acesso da população a serviços como hospitais, escolas, parques, eventos, entre outros. A dificuldade de acesso está relacionada principalmente com a qualidade do transporte público, tendo em vista que, em algumas cidades, o transporte por veículo próprio tem se tornado muito dispendioso a centros com grande fluxo de pessoas. A negligência de investimentos do poder público, culminou com o aumento do tempo de viagem entre pontos como, por exemplo, casa/trabalho/casa.

De tal forma desafiadora para mobilidade urbana, estudos tem apontado que as cidades estão crescendo de forma vertiginosa, estima-se que 50% da população do planeta está morando em centros urbanos e este valor chegará a 70% em 2050, segundo World Health Organization (2016). De acordo com o World Atlas (2018), há 200 anos, apenas as cidades de Londres, Pequim e Tóquio tinham mais de 1 milhão de habitantes e em 2018 135 cidades já chegaram a essa marca. Essas estimativas também vão ao encontro do relatório divulgado pela International Organization for Migration (2015), em que mais da metade da população global (54%) vivia em áreas urbanas, e projetou-se que 2,5 bilhões de habitantes urbanos serão adicionados até 2050.

Entender o deslocamento e buscar um padrão para identificar as necessidades do movimento das pessoas em uma região tem sido tema de trabalhos para pesquisadores desde o trabalho de Ravenstein (1885). No passado, isso era conseguido por meio da coleta de dados da pesquisa em pequenas amostras e baixas frequências (por exemplo, agências de planejamento de áreas metropolitanas nos países desenvolvidos realizam 1% das pesquisas de viagens domésticas uma ou duas vezes em uma década). Com a evolução da sociedade e a inovação em tecnologia, as cidades se tornaram mais diversificadas e complexas do que nunca e o mundo está cada vez mais interconectado.

No Brasil, a desigualdade social se reflete no acesso a serviços associados a mobilidade urbana como escolas, hospitais, áreas de lazer, entre outros. A constituição federal, no artigo 21, determina que a união institua "diretrizes para o desenvolvimento urbano, inclusive habitação, saneamento básico e transportes urbanos". O artigo 182, por sua vez, está prevendo que a política do desenvolvimento urbano executado pelo Poder Público

municipal deverá seguir as diretrizes fixadas na lei nº 12.587 para que se ordena "o pleno desenvolvimento das funções sociais da cidade e garantir o bem estar de seus habitantes" (SENADO, 1988), (PLANALTO, 2019).

Esse aumento populacional combinado com a necessidade de planejamento dos centros urbanos criou um campo com enorme potencial e é uma grande fonte de inspiração e estudo para planejadores urbanos e cientistas sociais, mas com o crescimento populacional das cidades, essas tarefas ficaram muito custosa e acabam por não atender aos desafios atuais, além de limitar os pesquisadores a trabalhar com amostras pequenas, se comparado ao fluxo de informações que se deseja analisar para obter uma melhor precisão.

Essa necessidade tem atraído pesquisadores da área da ciência de dados para o desenvolvimento de sistema que possam coletar informações de diversos dispositivos, armazenar uma grande massa de dados, manipular para que possam organizar a fim de extrair conhecimento sobre os dados. Isso gerou um crescimento no campo de pesquisas para melhorar o planejamento das cidades e a mobilidade urbana.

1.1 CONTEXTUALIZAÇÃO

Compreender os fluxos de tráfego em um ambiente urbano, estudar semelhanças (ou dissimilaridades) entre os dias da semana, encontrar os picos de trafego em um dia são exemplos de tarefas necessárias para entender a mobilidade urbana. Segundo Silva (2014), mobilidade urbana é definida como a facilidade de deslocamento das pessoas e bens na cidade, com o objetivo de desenvolver atividades econômicas e sociais no perímetro urbano de cidades, aglomerações urbanas e regiões metropolitanas. Tais deslocamentos são realizados por meio de veículos motorizados e não motorizados, elétricos, tração humana e até mesmo a pé, utilizando a infraestrutura disponibilizada nas regiões, dentre as quais ruas, avenidas, estradas, ciclovias e calçadas, que permitem o ir e vir cotidiano.

Para apoiar a análise do tráfego urbano, identificar as características e o perfil de atividades em diferentes áreas em uma cidade pode ser estratégico. Diferentes áreas podem ter características diferentes baseado em atividades como i)social, comercial, diversão e atividades turísticas, ii) pessoas e fluxo de trafego, iii) custo de vida, etc. A similaridade entre diferentes áreas pode ser extraída, mesmo quando áreas não estão próximas geograficamente. Como um exemplo, áreas caracterizadas pela presença de entretenimento ou gastronomia pode ser encontradas em centros e/ou áreas suburbanas. Além disso, estações de ônibus estão espalhadas por toda a cidade e grandes fluxos de pessoas cruzam frequentemente as áreas no qual as estações estão localizadas. (D'Andrea et al., 2018).

Diversas fonte de dados podem ser uteis para caracterizar atividade realizadas em uma área de uma cidade. Segundo Marjani et al. (2017), têm aumentado o número

de cidades que investem em soluções tecnológicas e dispositivos *IoT* como câmeras de monitoramento e sensores de presença, os quais estão cada vez mais presentes em ruas, praças, centro de compras, pontos turísticos, centros de eventos, no transporte público, etc. Além de dispositivos de uso pessoal com celulares, relógios e carros inteligentes.

Além do mais, a Internet, mais especificamente as redes sociais, têm produzido diariamente novos dados sobre atividades e o estilo de vida de seus usuário. *Posts* georreferenciados, avaliação do usuário sobre um serviço e comentários podem ser extraídos de redes sociais para extrair características significativas para descrever o perfil das áreas de uma cidade.

Por fim, para analisar a enorme e heterogênea quantidade de dados extraídos, aplicações que incorporem técnicas de mineração de dados tem sido adotadas frequentemente na literatura, vários trabalhos apresentaram abordagens para análise da mobilidade urbana (BATTY, 2009), (CLARAMUNT; JIANG; BARGIELA, 2000), (DEMISSIE; CORREIA; BENTO, 2013), (JIANG; FERREIRA; GONZÁLEZ, 2012) e para rotular regiões (ANDRIENKO et al., 2013), (SONG; MILLER, 2012), (D'Andrea et al., 2018), (JIANG; FERREIRA; GONZÁLEZ, 2012). Esses trabalhos usam técnicas de aprendizado de máquina (por exemplo, classificador e agrupamento) para lidar com dados adquiridos de dispositivos *IoT*, sensores, registro de chamadas de telefonia móvel, pontos de interesse, postagens online, informações de tráfego e outras fontes de dados. Para melhorar a análise do usuário, técnicas de visualização de informações são empregadas para aprimorar os layouts dos mapas e apresentar visualizações nos mapas da cidade. Embora categorizar as regiões da cidade com base em técnicas de agrupamento possa fornecer informações sobre os padrões da cidade, tal abordagem pode ocultar informações importantes devido à agregação.

1.2 *OBJETIVOS*

O objetivo deste trabalho foi desenvolver uma aplicação baseado na web chamado PlaceProfile para auxiliar na criação de perfis e caracterização de áreas de uma cidade com base na análise visual de pontos de interesse (POIs) de regiões de uma cidade. Para tanto, dado um conjunto de POIs e suas categorias, a abordagem proposta utiliza uma visualização iconográfica para criar perfis de áreas com base nas principais categorias de POIs, além de empregar técnicas de agrupamento para rotular áreas de uma região com base na frequência dos POIs presentes em cada área. A abordagem iconográfica amplia o poder da análise de agrupamento, mostrando como os POIs em diferentes áreas da cidade estão relacionados, bem como apresentando os detalhes das informações que podem ser usados para diferenciar os resultados do agrupamento. Adicionalmente, o PlaceProfile também apresenta uma estratégia de coordenação entre as áreas do mapa apoiada pelo

gráfico Radar Chart. Assim, esses mecanismos de coordenação também fornecem uma maneira de explicar os resultados do agrupamento para ajudar os analistas na análise detalhada, mostrando atividades de POI predominantes para cada área.

Os resultados obtidos por meio da análise realizada pelo PlaceProfile beneficiará uma variedade de aplicações, por exemplo, aquelas com o objetivo de apoiar governos locais para o planejamento urbano, fornecer informações aos turistas e às pessoas para poderem avaliar a escolha de casas para alugar ou hotéis baseado em torno das atividades características daquela região, para melhorar a tomada de decisão em negócios e publicidade, para donos de negócios poderem identificar a melhor área para abrir uma atividade comercial.

Resumidamente, as principais contribuições de PlaceProfile são:

- Uma abordagem de visualização iconográfica para mostrar rapidamente as principais atividades de áreas em uma cidade;
- Uma abordagem de coordenação entre as áreas de interesse e um gráfico de radar para ajudar na explicabilidade dos resultados de agrupamento e criação de perfil.

Assim, os dados produzidos pelo PlaceProfile será de grande utilidade para pesquisas que procuram entender o deslocamento urbano, assim melhorando a acurácia nos resultados da análise dessas pesquisas que tentam entender para quais finalidades grupos de pessoas se deslocam de um ponto a outro em uma cidade ou metrópole.

1.3 *ORGANIZAÇÃO*

Este documento está organizado da seguinte maneira.

- No Capítulo 1 foi contextualizado o problema na qual as cidades tem enfrentado com o crescente aumento populacional e como a tecnologia, aliado a ciência de dados, tem proposto soluções.
- No Capitulo 2 é apresentado a fundamentação téorica, onde são discutidos os principais conceitos utilizados neste trabalho.
- No Capítulo 3 descreve a revisão da literatura sobre pesquisas na área da mobilidade urbana, como foco em mineração dos dados e técnicas de visualização para representação dos dados.
- No Capítulo 4 é apresentado os métodos usados para o desenvolvimento do software, desde a coleta até a obtenção dos resultados finais.
- No Capítulo 5, descreve os resultados visuais obtidos através do uso do software em um conjunto de dados de amostra previamente coletado.

- Finalmente, na Seção 6 é feito uma comparação desta pesquisa com as demais que procuram rotular regiões, suas limitações e propostas futuras.

2 FUNDAMENTAÇÃO

2.1 EXTRAÇÃO DE CONHECIMENTO

Para extrair conhecimento a partir de dados coletados por diversos meios em uma cidade ou região, são necessárias várias etapas para se obter respostas a partir de um grande volume de dados. Assim, a descoberta de conhecimento procura relacionar dados para identificação de padrões compreensíveis, válidos, novos e potencialmente úteis. O processo para extrair conhecimento é representado de forma simplificada na Figura 1, o processo inicia pela coleta e armazenamento dos dados em sua forma bruta, em seguida os dados são pre-processados para que possam ser utilizados apenas os dados de interesse, os quais são minerados para que o usuário possa obter informações relevantes para um melhor entendimento e tomada de decisão.

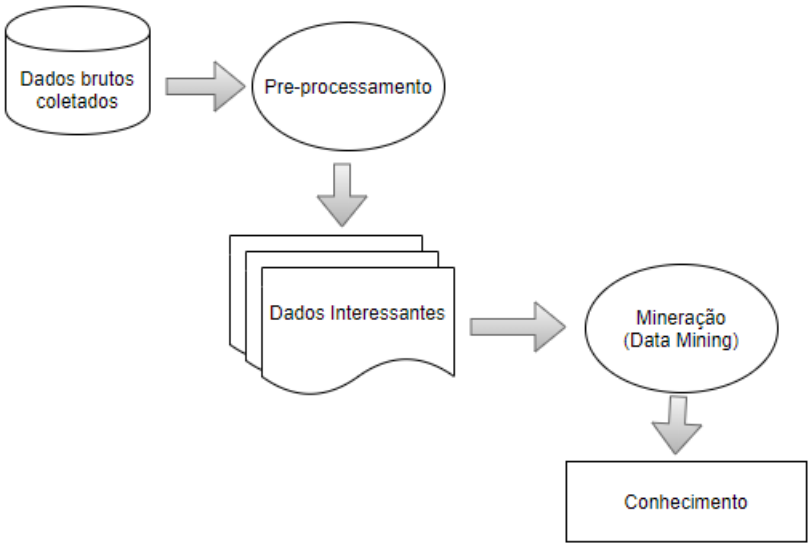

Figura 1 – Processo de extração de conhecimento inicia na obtenção de dados brutos coletados de diversos meios, em seguida é realizado o pre-processamento em que os dados são limpos e organizados para que possa ser utilizados apenas os dados de interesse. Os dados interessantes são enfim minerados para que o processo finalize e permita ao usuário obter informações relevantes.

2.1.1 OBTENÇÃO DE DADOS BRUTOS

Segundo Marjani et al. (2017), uma quantidade volumosa de dados está sendo produzida desde a década passada por diversos meios. O uso de dispositivos conhecidos como *IoT*(Internet das Coisas, do inglês *Internet of Things*), têm aumentado nas cidades

que investem em soluções tecnológicas, além disso diversos equipamentos eletrônicos de uso pessoal como eletrodomésticos, carros, celulares, câmeras, relógios, etc, têm saído de fábrica com micro *softwares* embutidos que permitem a integração com outros dispositivos. Esses dispositivos estão contribuindo significativamente para essa grande produção de dados. Contudo, sem um processo analítico, esses dados não são úteis para obtenção de valiosos *insights* em uma ampla pesquisa.

Dispositivo *IoT* tem surgido como a nova tendência nos últimos anos, no trabalho de Chen et al. (2014), estimou-se que em 2017 haveria mais de 50 bilhões de dispositivos coletando dados em tempo real e que em 2030 o mundo terá a quantidade de 1 trilhão de dispositivos móveis dos quais: relógios de pulso, máquinas de venda, alarmes, câmeras, instalações de transporte, instalações pública, eletrodomésticos, etc, estarão conectados e podem ser usados como equipamentos de aquisição de dados. Atualmente, esses dispositivos estão conectados à Internet por meio de redes de acesso que utilizam comunicação sem fio, para transmitir os dados e receber comandos variados, por meio de tecnologias como *Bluetooth*, *WiFi*, *ZigBee*, GPS, RFID e GSM sendo usadas para tal fim, permitindo a integração do mundo físico a sistemas de computação.

Outra interessante fonte de dados, que diariamente produz novos dados e em larga escala são oriundos de fontes online na web, mais especificamente de redes sociais. Tais dados possuem informações sobre locais de interesse na cidade, custo de vida, tráfego, preferência, etc. A maioria das fonte de dados na web, em geral, fornecem informações sobre qualquer cidade ou região geográfica, enquanto outras fontes de dados são específicas para uma determinada cidade (D'Andrea et al., 2018). Entre as fontes de dados mais conhecidas na web podemos citar os serviços do Facebook[1], Twitter[2] e Foursquare[3] que são úteis para recuperar informações relacionadas a rastros deixados pelos humanos, usuários destes serviços, enquanto os serviços do AirBnb[4], Booking[5], Tripadvisor[6] e Yelp[7] são exemplos de fonte de dados úteis na web que fornecem informações relacionadas ao estilo de vida de seus usuários. Tais fontes ainda permitem coletar informações sobre o valor do aluguel e características dos imóveis, úteis por exemplo para análise do custo de vida em uma determinada região. Para informações detalhadas sobre atividades de negócios na web, o Google Maps[8] fornece informações sobre pontos de interesse (POIs)divulgados pelos próprios usuários no mapa da cidade, uma espécie de páginas amarelas onde os dados podem ser facilmente coletados usando a API *Google Place*. Para cada POI coletado é possível obter informações exatas sobre a posição em coordenadas de latitude e longitude,

[1] https://www.facebook.com/
[2] https://twitter.com/
[3] https://foursquare.com/
[4] https://www.airbnb.com.br/
[5] https://www.booking.com/
[6] https://www.tripadvisor.com/
[7] https://www.yelp.com/
[8] https://www.google.com/maps

nome do local, endereço, categoria (museu, estação ferroviária, bar, escola, loja, ponto de ônibus, supermercado, etc.). Como alternativa ao Google Maps, o projeto OpenStreetMaps[9] é livre e tem um protocolo aberto para que todos os utilizadores possam ajudar a melhorá-lo. É desenvolvido por uma comunidade voluntária de utilizadores que vão atualizando os dados sobre endereços, estradas, nomes de lugares e planejamento de rotas.

2.1.2 MINERAÇÃO DOS DADOS

Segundo Marjani et al. (2017), a mineração é o processo de examinar um grande conjunto de dados que contém uma variedade de tipos de dados para revelar padrões despercebidos, correlações escondidas, tendências de mercado, preferência dos clientes e poderosas informações sobre negócios. A capacidade de analisar grande quantidade de dados pode ajudar uma organização a lidar com informações que podem afetar os negócios. Portanto, o principal objetivo é apoiar no entendimento dos dados e, assim, tomar decisões eficientes.

Métodos de mineração são amplamente usados para problemas de análise de grandes volumes de dados como métodos estatísticos e de *machine learning* (ML), os quais são excelentes meios de se extrair valor desses dados a um custo computacional baixo (MUKHOPADHYAY et al., 2013). No trabalho de Marjani et al. (2017) os autores analisaram 4 categorias de métodos de análise de dados: classificação, agrupamento, regras de associação e predição. Cada categoria classifica um grupo de algoritmos de mineração de dados para atender aos requisitos de extração e análise de informações. A Figura 2, descreve e resume cada uma dessas categorias analisadas. Cada categoria é uma função de mineração de dados e envolve muitos métodos e algoritmos para atender aos requisitos de extração e análise de informações.

A classificação é uma abordagem de aprendizado supervisionado que utiliza base de dados com observações já conhecidas como dados de treinamento para classificar objetos de dados em grupos que ainda não foram rotulados (ESTIVILL-CASTRO, 2002). Uma categoria predefinida é atribuída a um objeto e, portanto, o objetivo de prever um grupo ou classe para um objeto é alcançado (ver Figura 2). Entre os algoritmos de classificação mais populares temos as Redes Bayesianas (*Bayesian Network*) (BIELZA; LARRAÑAGA, 2014), *Support Vector Machine* (SVM) (SUYKENS; VANDEWALLE, 1999) e o *k-nearest neighbor* (KNN) (KELLER; GRAY; GIVENS, 1985). As Redes Bayesianas são eficientes para analisar estruturas de dados complexas, em vez dos formatos de dados estruturados tradicionais (BIELZA; LARRAÑAGA, 2014). A análise de padrões de dados e a criação de grupos são executadas com eficiência usando o SVM, que também é uma abordagem de classificação. O SVM utiliza a teoria estatística do aprendizado para analisar padrões de dados e criar grupos. Aplicações da classificação SVM na análise de dados

[9] https://www.openstreetmap.org/

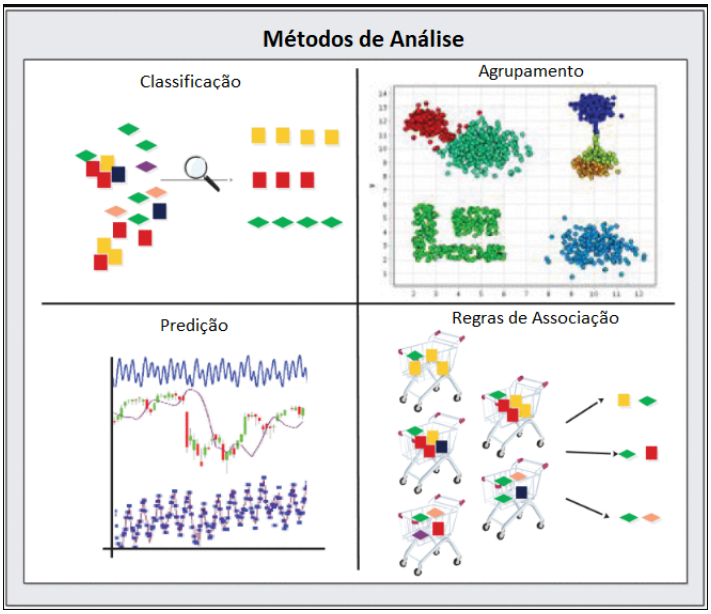

Figura 2 – Uma visão Geral dos métodos analisados no trabalho de Marjani et al. (2017).

incluem classificação de texto, correspondência de padrões, diagnóstico de saúde e comércio (SUYKENS; VANDEWALLE, 1999). Da mesma forma, o KNN é normalmente projetado para fornecer mecanismos eficientes para encontrar padrões ocultos a partir de grandes conjuntos de dados, de modo que os objetos recuperados sejam semelhantes à categoria predefinida (KELLER; GRAY; GIVENS, 1985).

O agrupamento (ou *Clustering*, em inglês) é outra técnica de mineração usada como método analítico de dados. Ao contrário da classificação, o agrupamento usa uma abordagem de aprendizado não supervisionada e cria grupos para determinados objetos com base em suas características significativas e distintas (SRIVASTAVA et al., 2013) (ver Figura 2). Aplicar técnicas de agrupamento em um grande número de objetos na forma de *clusters* (grupos), auxilia na simplificação e facilita a manipulação de dados. Os métodos de agrupamento se dividem em duas outras categorias, os métodos de particionamento, como o algoritmo *K-means* (KRISHNA; MURTY, 1999), *Fuzzy C-Means* (CANNON; DAVE; BEZDEK, 1986) e *Agglomerative* (KURITA, 1991), e os métodos hierárquicos como o algoritmo DBSCAN (SCHUBERT et al., 2017).

Segundo Gosain e Bhugra (2013), o processo de mineração de dados por meio do algorítimo de regras de associação envolve a identificação de relacionamentos interessantes entre diferentes objetos, eventos ou outras entidades para analisar tendências de mercado, comportamento de compra do consumidor e previsões de demanda de produtos (ver 2). O algoritmo de regras de associação concentra-se na identificação e criação de regras com base na frequência de ocorrências para dados numéricos e não numéricos. O processamento de dados é realizado de duas maneiras, de acordo com as regras de associação. Primeiro, o processamento sequencial de dados usa o algoritmo baseado em priori, como MSPS

Método	**Aplicações**										
	Gerenciamento de desastres	Assistência médica	Imagem clínica	Genética	Análise de mercado	Indústria	Reconhecimento de fala	Bioinformática	Processamento de linguagem natural	Análise de rede social	e-govenança
Classificação	-	-	x	-	-	x	x	-	x	-	x
Agrupamento	-	x	x	x	x	x	-	x	-	x	x
Regras de associação	-	x	-	-	x	x	-	x	-	-	x
Predição	x	-	-	-	x	-	-	-	-	x	-
Séries temporais	x	-	x	-	-	-	x	-	-	x	x

Tabela 1 – Matriz que relaciona os métodos de mineração de dados para aplicações em diversas áreas, o 'x' indica que a técnica de mineração pode ser usar para as respectivas aplicação, enquanto o '-' não é óbvio que a técnica pode ser usada. (Marjani et al., 2017).

(LUO; CHUNG, 2005) e LAPIN-SPAM (YANG; KITSUREGAWA, 2005), para identificar associação de interação. Uma outra abordagem é o processamento de dados é a análise de sequência temporal (HATEREN; RUDERMAN, 1998), que usa algoritmo para analisar padrões de eventos em dados contínuos.

A análise preditiva usa dados históricos, conhecidos como dados de treinamento, para determinar os resultados como tendências ou comportamento nos dados. Os algoritmos SVM e lógica fuzzy são usados para identificar relações entre variáveis independentes e dependentes e obter curvas de regressão para previsões, como para desastres naturais (GANDOMI; HAIDER, 2015). Além disso, as previsões de compra do cliente e as tendências de mídia social são analisadas por meio de análises preditivas. Os requisitos de processamento são modificados de acordo com a natureza e o volume dos dados. Métodos de acesso e mineração rápidos para dados estruturados e não estruturados são as principais preocupações relacionadas à análise de dados. Além disso, a representação de dados é um requisito significativo. A análise de séries temporais reduz a alta dimensionalidade associada a grandes volumes de dados e oferece representação para melhor tomada de decisão.

Os métodos de análise de *big data* são amplamente adotados em muitas áreas de aplicação de big data, como gerenciamento de desastres, assistência médica, negócios,

setor e governança eletrônica. Na Tabela 1, Marjani et al. (2017) apresenta as áreas de aplicação da mineração de *big data* que foram discutidos nesta seção, o 'x' indica que a técnica é usada para as aplicações, enquanto '-' indica que não é claro se o método apoia ou não as aplicações. Em particular, a Tabela 1 mostra que os métodos de classificação são adequados para imagens médicas, indústria, reconhecimento de fala, processamento de linguagem natural e governança eletrônica. Os métodos de análise de dados baseados em regras de cluster e associação são aplicáveis ao setor e à governança eletrônica e são bem adotados em assistência médica, comércio eletrônico e bioinformática. A análise preditiva é útil para previsões de desastres e de mercado, enquanto a análise de séries temporais é usada na previsão de desastres, imagens médicas, reconhecimento de fala, análise de redes sociais e governança eletrônica.

2.2 ALGORITMOS DE AGRUPAMENTO

Como já discutido na Seção 2.1.2, os algoritmos de agrupamento (ou *Clustering*, em inglês) são usados como métodos analíticos de dados. Eles usam uma abordagem de aprendizado não supervisionada e ter por objetivo criar grupos para determinados objetos com base em suas características significativas e distintas.

Os algoritmos de agrupamento classificam os dados brutos e busca os padrões ocultos que podem existir nos conjuntos de dados (HUANG, 1998). É um processo de agrupar objetos de dados em clusters separados, de forma que os dados no mesmo cluster sejam semelhantes de acordo com os atributos que descrevem os dados e as técnicas empregadas para gerar os clusters. A demanda por organizar os dados cada vez maiores e aprender informações valiosas a partir dos dados, o que faz com que as técnicas de agrupamento sejam amplamente aplicadas em muitas áreas de aplicação, como inteligência artificial, biologia, gerenciamento de relacionamento com o cliente, compressão de dados, mineração de dados, recuperação de informações, processamento de imagem, máquina aprendizagem, marketing, medicina, reconhecimento de padrões, psicologia, estatística (Marjani et al., 2017), dentre muitas outras.

A seguir são apresentados detalhes de três algoritmos de agrupamento empregados neste trabalho.

2.2.1 *K-MEANS*

O algoritmo K-means foi proposto em 1967 no trabalho de MacQueen et al. (1967). É um método numérico, não supervisionado, não determinístico e iterativo. É simples e muito rápido, portanto, em muitas aplicações práticas, o método provou ser uma forma muito eficaz de produzir bons resultados de agrupamento.

No trabalho de Na, Xumin e Yong (2010), os autores discutiram o algoritmo K-means e propor melhorias. Ainda, os autores detalharam o funcionamento do algoritmo que consiste em duas fases distintas. A primeira fase seleciona k centros aleatoriamente, onde o valor k é previamente fixado. A próxima fase é levar cada objeto de dados ao centro mais próximo. A distância euclidiana é geralmente considerada para determinar a distância entre cada objeto de dados e os centros do cluster. Quando todos os objetos de dados são incluídos em alguns clusters, a primeira etapa é concluída e um agrupamento inicial é feito. Recalcular a média dos clusters formados anteriormente. Esse processo iterativo continua repetidamente até que a função de critério se torne o mínimo.

Supondo que o objeto de destino seja x, x_i indica a média do cluster C_i, a função de critério é definida da seguinte forma:

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} |x - x_i|^2 \tag{2.1}$$

E é a soma do erro quadrático de todos os objetos no banco de dados. A distância da função critério é a distância euclidiana, que é usada para determinar a distância mais próxima entre cada objeto de dados e o centro do cluster. A distância euclidiana entre um vetor $x = (x_1,x_2,\ldots x_n)$ e um outro vetor $y = (y_1,y_2,\ldots y_n)$, a distância euclidiana de $d(x_i,y_i)$, pode ser obtida da seguinte forma:

$$d(x_i, y_i) = \left[\sum_{i=1}^{n} (x_i - y_i)^2 \right]^{1/2} \tag{2.2}$$

O processo do algoritmo K-means segue como:

Entrada: Número de clusters desejados, k, e um banco de dados $D = d_1,d_2,..., d_n$, contendo n objetos de dados.

Resultado: Um conjunto de k clusters

Passos:

1. Selecione aleatoriamente k objetos de dados do conjunto de dados D como centros de cluster iniciais.
2. Repetir;
3. Calcule a distância entre cada objeto de dados d_i (1 <= i <= n) e todos os k centros de cluster c_j(1<=j<=k) e atribuir o objeto de dados d_i mais próximo do cluster
4. Para cada cluster j(1 < = j < = k), recalcule o centro do cluster.
5. Até que não exista mudança no centro dos clusters.

O algoritmo de agrupamento K-means sempre converge para o mínimo local. Antes de o algoritmo K-means convergir, cálculos de distância e centros de cluster são feitos enquanto as iterações são executados várias vezes, onde o inteiro positivo t é conhecido como o número de iterações K-means. O valor preciso de t varia dependendo dos centros de cluster iniciais (HUANG, 1998). A distribuição dos pontos de dados tem uma relação com o novo centro de agrupamento, então a complexidade do tempo computacional do algoritmo K-means é O(nkt). n é o número de todos os objetos de dados, k é o número de clusters, t é as iterações do algoritmo. Normalmente requer k « n e t « n.

2.2.2 *FUZZY C-MEANS*

Segundo Zhou e Ren (2019), um dos algoritmos de agrupamento mais amplamente usados é o algoritmo de agrupamento Fuzzy C-means (FCM). O algoritmo de agrupamento Fuzzy C-means (FCM) foi desenvolvido por JC Dunn em 1973 (DUNN, 1973), e melhorado por Bezdek (PEIZHUANG, 1983) em 1981. Atualmente, o FCM tornou-se um método de agrupamento bem conhecido e amplamente utilizado e é muito semelhante ao algoritmo K-means Zhou e Ren (2019).

O índice de desempenho que orienta o espaço de dados assume a seguinte forma:

$$Q = \sum_{i=1}^{c} \sum_{k=1}^{N} u_{ik}^{m} \left\| \mathbf{X}_k - \mathbf{V}_i \right\|^2 \tag{2.3}$$

As derivações da solução são concluídas em duas etapas. O primeiro envolve as restrições que acompanham os requisitos impostos à matriz de partição. Incorporamos as restrições com o auxílio de multiplicadores de Lagrange. Então, para cada padrão t = 1 , 2 , ... , N, nós formulamos o funcional aumentado:

$$V = \sum_{i=1}^{c} u_{it}^{m} d_{it}^{2} - \lambda \left(\sum_{i=1}^{c} u_{it} - 1 \right) \tag{2.4}$$

Onde λ denotam um multiplicador de Lagrange. Calculando a derivada de V em relação a u_{st} e tornando-o igual a 0, obtemos:

$$\frac{\partial V}{\partial u_{st}} = m u_{st}^{m-1} d_{st}^{2} - \lambda = 0 \tag{2.5}$$

e

$$u_{st} = \left(\frac{\lambda}{m} \right)^{1/(m-1)} \frac{1}{(d_{st})^{\frac{2}{m-1}}} \tag{2.6}$$

Levando em consideração a restrição de identidade $\sum_{j=1}^{c} u_{jt} = 1$, temos

$$\left(\frac{\lambda}{m}\right)^{1/(m-1)} \sum_{j=1}^{c} \frac{1}{(d_{jt})^{\frac{2}{m-1}}} = 1 \tag{2.7}$$

Isso nos permite determinar o multiplicador de Lagrange λ:

$$\left(\frac{\lambda}{m}\right)^{1/(m-1)} = \frac{1}{\sum_{j=1}^{c} \frac{1}{(d_{jt})^{\frac{2}{m-1}}}} \tag{2.8}$$

Em seguida, inserimos a expressão acima na Eq. 2.6, o que produz

$$u_{st}^{=} \frac{1}{\sum_{j=1}^{c} \left(\frac{d_{st}}{d_{jt}}\right)^{\frac{2}{m-1}}} \tag{2.9}$$

Os cálculos dos protótipos são diretos, pois nenhuma restrição é imposta a eles. O mínimo de Q calculado em relação a V_s rendimentos

$$\nabla_{\mathbf{v}_s} Q = 0 \tag{2.10}$$

A solução detalhada depende da função de distância. No caso da distância euclidiana, isso leva à expressão

$$2 \sum_{k=1}^{N} u_{sk}^{m} (\mathbf{X}_k - \mathbf{V}_s) = 0 \tag{2.11}$$

Nós imediatamente obtemos

$$\mathbf{V}_s = \frac{\sum_{k=1}^{N} u_{sk}^{m} \mathbf{X}_k}{\sum_{k=1}^{N} u_{sk}^{m}} \tag{2.12}$$

Para resumir, podemos considerar o algoritmo FCM como um processo iterativo envolvendo computações (atualizações) sucessivas dos protótipos e da matriz de partição. Os valores dos parâmetros são configurados antecipadamente. Eles consistem nos seguintes itens: o número de clusters (c), a função de distância $||\ \ldots||$, o fator de fuzzificação (m) e o critério de terminação (ϵ).

2.2.3 *AGGLOMERATIVE*

Segundo Rokach e Maimon (2005), o algoritmo padrão para agrupamento aglomerativo hierárquico, no inglês chamado de *hierarchical agglomerative clustering* (HAC),

tem uma complexidade de tempo de $\mathcal{O}(n^3)$ e requer $\Omega(n^2)$ memória, o que o torna muito lento até mesmo para conjuntos de dados médios. No entanto, para alguns casos especiais, métodos aglomerativos eficientes de complexidade $\mathcal{O}(n^2)$ são conhecidos: SLINK (SIBSON, 1973) para ligação única e CLINK (DEFAYS, 1977) para agrupamento de ligação completa. Em uma estrutura de dados baseado em árvore, o tempo de execução do caso geral pode ser reduzido para $\mathcal{O}(n^2 \log n)$, uma melhoria no limite acima mencionado de $\mathcal{O}(n^3)$, ao custo de aumentar ainda mais os requisitos de memória. Em muitos casos, a sobrecarga de memória dessa abordagem é muito grande para torná-la utilizável na prática.

Ainda, segundo Rokach e Maimon (2005), exceto para o caso especial de ligação única, nenhum dos algoritmos (exceto pesquisa exaustiva em $\mathcal{O}(2^n)$ pode ser garantido que encontrará a solução ideal.

O agrupamento divisivo com uma pesquisa exaustiva é $\mathcal{O}(2^n)$, mas é comum usar heurísticas mais rápidas para escolher divisões, como K-means.

2.3 VISUALIZAÇÃO DA INFORMAÇÃO

Segundo Tran e Le (2020), o ser humano conhece o mundo real ao perceber e analisar dados por meio de cinco órgãos dos sentidos, cada um dos quais coleta dados por um sensor. As orelhas coletam dados acústicos, narizes coletam dados de cheiro, língua coleta dados de sabor, pele coleta dados de toque e olhos coletam dados de imagem. Nesse sentido, os órgãos de visão que coletam dados pelos olhos é o canal mais amplo que auxilia o ser humano a conhecer e compreender o mundo real. Em outras palavras, os órgãos da visão são o canal mais importante para coletar e compreender imagens ou padrões em forma de dados.

A percepção da visão é um progresso criado pelos órgãos da visão humana para conhecer e compreender o que ele vê. Quando o ser humano olha para alguma coisa, os raios de luz que transportam dados da coisa são transmitidos a seus órgãos de visão para detectar informações. Os raios de luz passam pela retina e pela fóvea dos olhos, e então estes são decodificados no cérebro. As características da percepção da visão resultam na capacidade humana de compreender o que vê. Em outras palavras, é necessário pesquisar as características da percepção da visão para saber como constituir as melhores imagens ou padrões que representam dados no sistema de visualização (O'CONNOR, 2015).

A partir das definições acima, podemos dizer que, em um sistema de visualização de dados, os dados são convertidos em informações e/ou conhecimento pela percepção da visão humana do gráfico que representa esses dados. Segundo Tran e Le (2020), a eficácia de um padrão que representa os dados é avaliada pela taxa humana que percebe sua importância. A análise das características dos órgãos da visão humana na percepção de padrões resulta na constituição que permiti que os designers estruturem padrões que representem dados

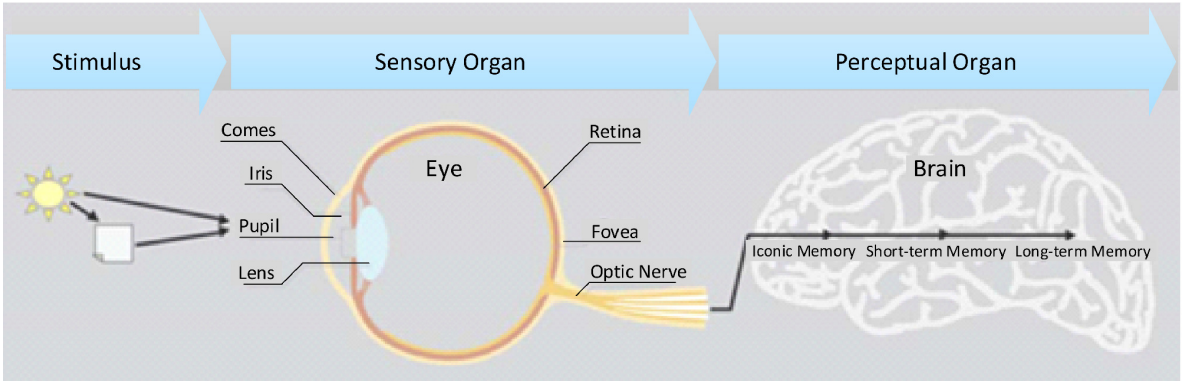

Figura 3 – Elementos no processo da interpretação da visão humana (O'CONNOR, 2015).

visualmente, de forma que o ser humano seja capaz de perceber muito do significado dos dados. A taxa efetiva de percepção de um padrão, por exemplo, um gráfico que representa os dados, que é formado pela aplicação das leis da Gestalt (HOCHBERG, 1957), é estimada objetivamente pelos seguintes recursos visuais:

- Associação - O recurso associativo refere-se a um padrão que permite ao ser humano perceber seus elementos semelhantes para agrupá-los em um *cluster*.
- Seleção - A característica seletiva se refere a um padrão que permite aos humanos perceber os elementos dominantes no padrão e distinguir grupos de elementos no padrão, cujas características não são semelhantes.
- Ordem - A característica ordinal se refere a um padrão que permite ao ser humano perceber a ordem dos elementos.
- Quantidade - O recurso quantitativo refere-se a um padrão que permite ao ser humano perceber o tamanho dos elementos, o intervalo ou a proporção do tamanho de dois elementos.
- Intervalo de valores - A faixa de valores refere-se à capacidade da tela de exibir todos os valores das variáveis.

A ideia básica da tecnologia de visualização de dados é que para cada item do banco de dados seja representado como um elemento visual, então um grande número de conjuntos de dados constituem a imagem dos dados (KEIM, 2002). Enquanto isso, usando dados multidimensionais para representar cada valor de atributo dos dados, os dados podem ser observados de diferentes dimensões e usados observações e análises mais aprofundadas. O principal objetivo da visualização de dados é transmitir informações usando ferramentas gráficas e se comunicar de forma clara e eficaz (BAO; CHEN, 2014).

Segundo Sobral, Galvão e Borges (2019), devido aos diferentes graus de dados, a visualização de dados deve implementar o recurso de zoom. Ao mesmo tempo, os usuários

podem navegar ou ter conhecimento específico sobre o conjunto de dados usando os gráficos de resposta dinâmica. As tecnologias padrão da Web para visualização de dados permitem que o usuário visualize as estatísticas em diferentes sistemas operacionais usando as versões mais recentes de navegadores.

2.3.1 TÉCNICAS DE VISUALIZAÇÃO

Segundo Keim (2000), as técnicas de visualização são cada vez mais importantes na exploração e análise de grandes quantidades de informações multidimensionais. A ideia básica da exploração de dados visuais é apresentar os dados por meio de representações gráficas, permitindo que o ser humano tenha uma visão dos dados, tire conclusões e interaja diretamente com os dados. Ainda, em seu outro trabalho, Keim (2002), afirmou que as técnicas de mineração visual de dados provaram ser de alto valor na análise exploratória de dados e também têm alto potencial para explorar grandes bancos de dados. A exploração visual de dados é especialmente útil quando pouco se sabe sobre os dados e os objetivos da exploração são vagos. Uma vez que o usuário está diretamente envolvido no processo de exploração, a mudança e o ajuste dos objetivos de exploração são feitos manualmente, se necessário.

Um outra afirmação de Keim (2002) é que há um grande número de técnicas de visualização que podem ser usadas para criar representações visuais de dados. Além das técnicas 2D/3D padrão, como plotagens x-y (x-y-z), gráficos de barras, gráficos de linha, etc., existem várias técnicas de visualização mais sofisticadas. Tais técnicas correspondem a princípios básicos de visualização que podem ser combinados para implementar um sistema de visualização específico.

2.3.2 TÉCNICAS ORIENTADAS A PIXEL

Uma classe importante de técnicas de visualização que é particularmente interessante para visualizar conjuntos de dados multidimensionais muito grandes é a classe de técnicas orientadas a pixels. A ideia básica das técnicas de visualização orientada a pixels é representar o maior número possível de objetos de dados na tela ao mesmo tempo, mapeando cada valor de dados para um pixel da tela e organizando os pixels de forma adequada, veja na Figura 4.

Todas as técnicas orientadas a pixels particionam a tela em várias subjanelas. Para conjuntos de dados com m dimensões (atributos), a tela é particionada em m subjanelas - uma para cada uma das dimensões. No caso de uma classe especial de técnicas orientadas para pixels - as técnicas dependentes de consulta - uma janela adicional (m + 1) é fornecida para a distância total. Dentro das janelas, os valores dos dados são organizados de acordo com a classificação geral fornecida, que pode ser orientada por dados para as

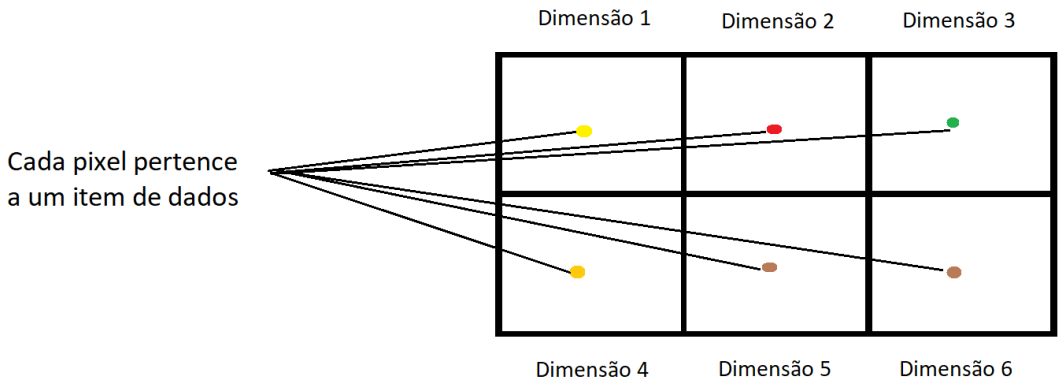

Figura 4 – Arranjo básico de subjanelas de dados com seis dimensões, adaptado para o português do trabalho de Keim (2000).

técnicas independentes de consulta ou orientada por consulta para as técnicas dependentes de consulta. Correlações, dependências funcionais e outras relações interessantes entre dimensões podem ser detectadas relacionando regiões correspondentes nas múltiplas janelas. Em geral, tais técnicas permitem visualizar a maior quantidade de dados possível em exibições atuais (até cerca de 1.000.000 de valores de dados) (KEIM, 2000).

Para atingir esse objetivo, uma série de questões de design devem ser resolvidas. Keim (2000), em seu trabalho, relacionou questões importante a serem resolvidas em uma visualização orientada a pixel.

- O primeiro problema é o mapeamento de valores de dados para cores. Um bom mapeamento é obviamente muito importante, mas deve ser cuidadosamente projetado para ser intuitivo.
- A segunda questão muito importante é como os pixels são organizados dentro das subjanelas. O arranjo depende dos dados e da tarefa de visualização e, portanto, arranjos diferentes são úteis para finalidades diferentes. O problema do arranjo pode ser descrito formalmente como uma questão de otimização e diferentes técnicas de visualização otimizam diferentes variantes do problema de otimização.
- Uma terceira questão é a forma das subjanelas. Quando as subjanelas possuem a forma retangular, para os conjuntos de dados com um grande número de dimensões (atributos), as subjanelas para as diferentes dimensões ficam bastante distantes e, portanto, torna-se difícil encontrar relações interessantes entre as dimensões. Como solução para este problema, os autores propuseram encontrar formas alternativas para as subjanelas da dimensão.
- A próxima questão ao projetar as técnicas orientadas a pixels é como ordenar as subjanelas para as dimensões (atributos). Na maioria das aplicações, não há

ordenação natural das dimensões. Para detectar dependências e correlações entre as dimensões representadas nas subjanelas, é melhor colocar dimensões relacionadas próximas umas das outras.

2.3.3 TÉCNICAS ICONOGRÁFICAS

Outra classe de técnicas de exploração de dados visuais são as técnicas de exibição iconográficas. A ideia é mapear os valores de atributo de um item de dados multidimensional para os recursos de um ícone. Os ícones podem ser definidos arbitrariamente, veja na Figura 5, eles podem ser (a)ícones de agulha, (b) rostos, ícones de bonecos de palito (c) , ícones de estrela (d), ícones coloridos (e), mapas de grades (f). A visualização é gerada mapeando os valores de atributo de cada registro de dados para os recursos dos ícones. No caso da técnica do boneco palito, por exemplo, duas dimensões são mapeadas para as dimensões do display e as dimensões restantes são mapeadas para os ângulos e/ou comprimento do membro do ícone do boneco palito. Se os itens de dados forem relativamente densos em relação às duas dimensões de exibição, a visualização resultante apresenta padrões de textura que variam de acordo com as características dos dados e são, portanto, detectáveis pela percepção pré-atencional (KEIM, 2002).

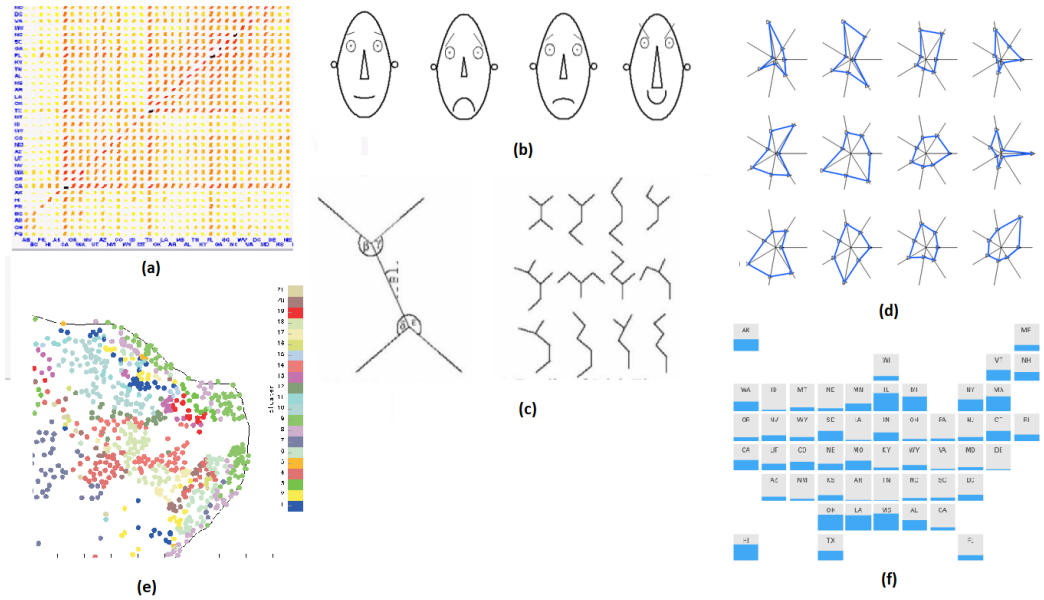

Figura 5 – Exemplos de técnicas de visualização iconográfica. (a) ícones de agulha (WARD, 1994), (b) rostos (ASTEL et al., 2006), (c) ícones de bonecos de palito (KAISER, 2000), (d) ícones de estrela (JMP, 2021), (e) ícones coloridos (NOCKE; SCHLECHTWEG; SCHUMANN, 2005), (f) Mapas de grades (DATAVIZCATALOGUE, 2021).

2.3.4 TÉCNICAS HIERÁRQUICAS

As técnicas de display empilhada ou técnicas hierárquicas são personalizadas para apresentar os dados particionados de maneira hierárquica. No caso de dados multidimensionais, as dimensões de dados a serem usadas para particionar os dados e construir a hierarquia devem ser selecionadas apropriadamente. A ideia básica é incorporar um sistema de coordenadas dentro de outro sistema de coordenadas, ou seja, dois atributos formam o sistema de coordenadas externo, dois outros atributos são incorporados ao sistema de coordenadas externo e assim por diante. A exibição é gerada dividindo os sistemas de coordenadas de nível mais externo em células retangulares e, dentro das células, os próximos dois atributos são usados para abranger o sistema de coordenadas de segundo nível. Este processo pode ser repetido mais uma vez. A utilidade da visualização resultante depende em grande parte da distribuição de dados das coordenadas externas e, portanto, as dimensões que são usadas para definir o sistema de coordenadas externas devem ser selecionadas com cuidado. A regra é escolher primeiro as dimensões mais importantes (veja na Figura 6). Outros exemplos de técnicas de exibição empilhadas incluem Worlds-within-Worlds, Treemap e Cone Trees (KEIM, 2002).

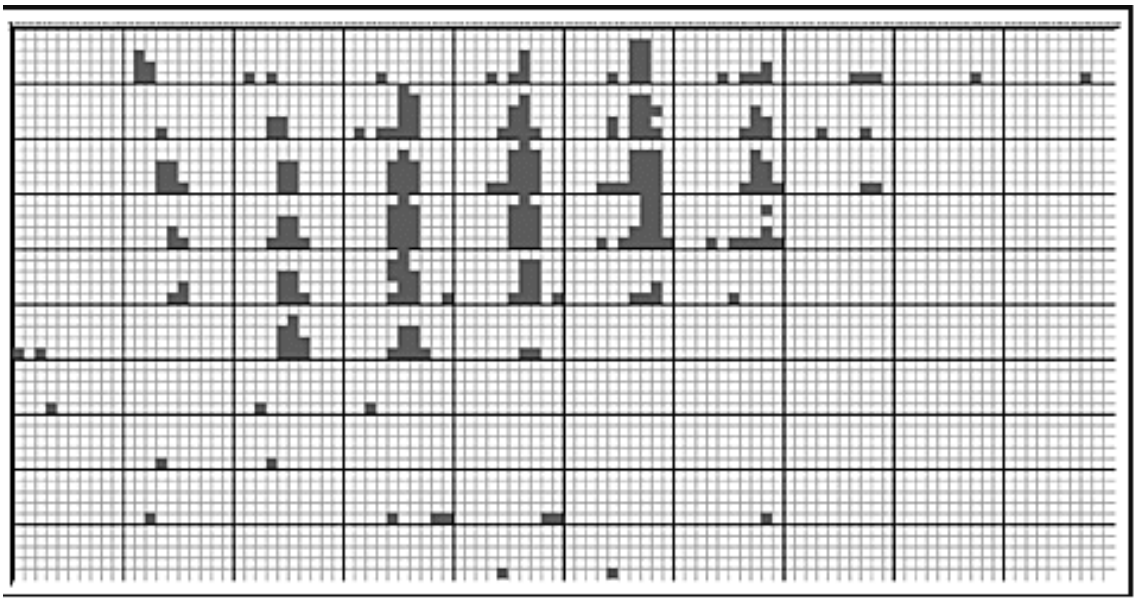

Figura 6 – Visualização da técnica de hierarquia dimensional de dados (KEIM, 2002).

2.3.5 PROJEÇÕES GEOMÉTRICAS

Segundo Keim (2002), as técnicas de exibição transformadas geometricamente visam encontrar transformações interessantes de conjuntos de dados multidimensionais. A classe de técnicas de exibição geométrica inclui técnicas de estatística exploratória, como matrizes de gráfico de dispersão e técnicas que podem ser incluídas no termo busca de projeção. Outras técnicas de projeção geométrica incluem Prosection Views, Hyperslice e

a conhecida técnica de visualização de Coordenadas Paralelas. A técnica de Coordenadas Paralelas mapeia o espaço k-dimensional nas duas dimensões de exibição usando k eixos equidistantes que são paralelos a um dos eixos da tela. Os eixos correspondem às dimensões e são escalonados linearmente do valor mínimo ao máximo da dimensão correspondente. Cada item de dados é apresentado como uma linha poligonal, intersectando cada um dos eixos naquele ponto que corresponde ao valor das dimensões consideradas (Veja na Figura 7).

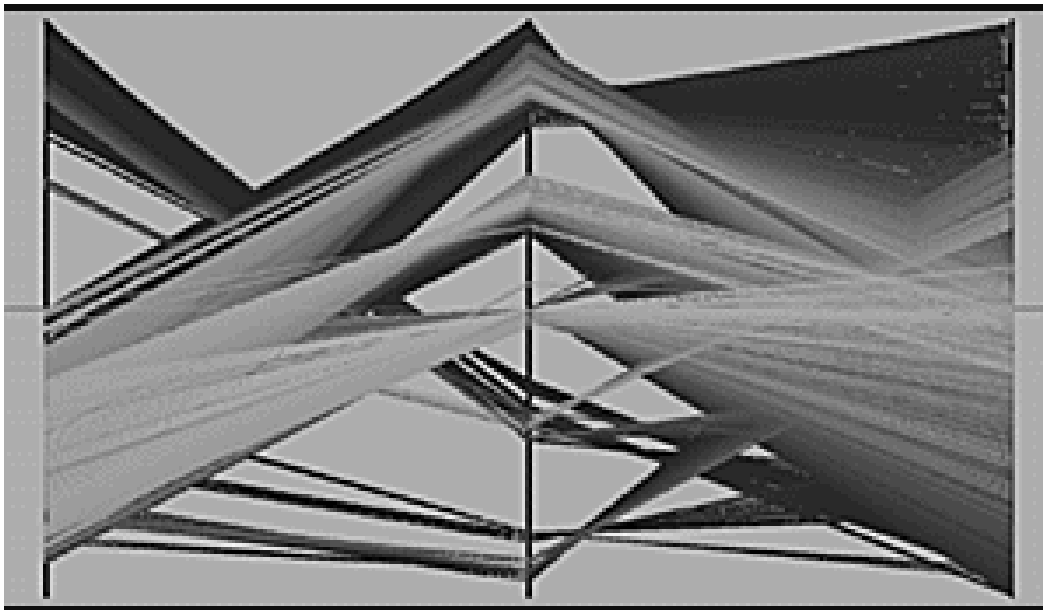

Figura 7 – Visualização da técnica de Coordenadas Paralelas (KEIM, 2002).

3 TRABALHOS RELACIONADOS

As primeiras aplicações dos sistemas de visualização para análise da mobilidade urbana baseavam-se nos recursos dos Sistemas de Informação Geográfica (SIG) e nos métodos tradicionais de visualização, por exemplo, gráficos de barras, de linhas e representações baseadas em mapas na forma de *heatmaps* (mapas de calor) ou coropletas, com recursos de interação limitados (CLARAMUNT; JIANG; BARGIELA, 2000). Um SIG permitia a visualização em diferentes perspectivas. No exemplo apresentado na Figura 8, a aplicação proposta no trabalho de Claramunt, Jiang e Bargiela (2000), permite ao usuário analisar o monitoramento do tráfego urbano em quatro perspectivas: (a) análise dados espaciais com o uso de mapas geovisuais, (b) gráficos de áreas para identificar regiões de maior fluxo, (c) série temporal, em um gráfico de linhas para fazer a análise em intervalos de datas e (d) gráfico de barras para análise do intervalo entre valores. Apresentar a visualização em diferentes níveis de agregação forneceu uma boa referência para trabalhos futuros sobre como diferentes visualizações podem ser úteis para especialistas em um domínio e em diferentes contextos.

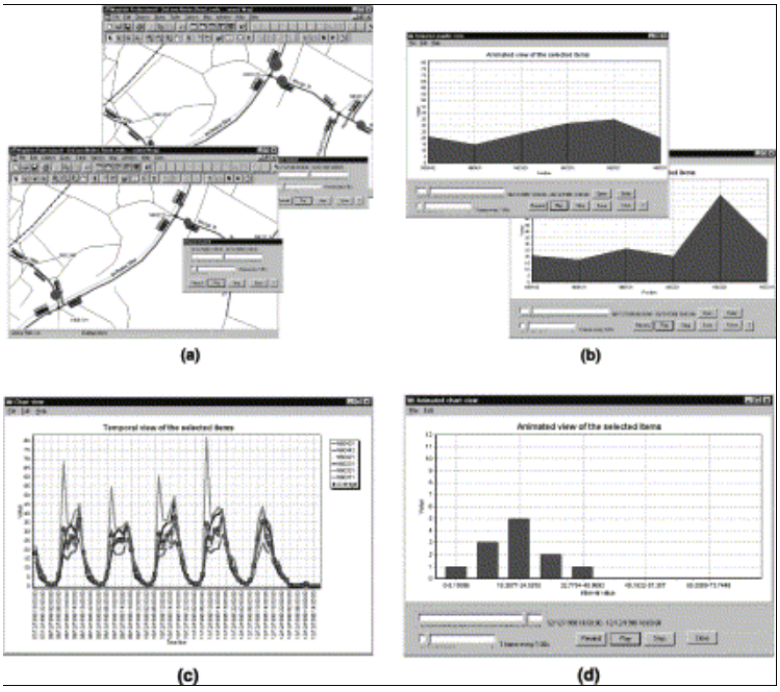

Figura 8 – Um dos primeiros trabalhos sobre monitoramento do tráfego urbano, na Figura (a) os fluxos de dados são analisados espacialmente, em (b) tematicamente com gráficos de áreas, em (c) temporalmente, através de gráficos de linhas que representam séries temporais e (d) com agregação, usando gráfico de barras para mostrar intervalo entre valores (CLARAMUNT; JIANG; BARGIELA, 2000).

Nesta década, os avanços na computação gráfica e suas tecnologias produziram

novos arcabouços de visualização, por exemplo, D3.js[1], Google Maps API[2], Mapbox[3], Carto[4], ArcGIS[5], entre outros. Tais arcabouços são amplamente utilizadas na literatura para a descoberta de conhecimento de dados produzidos pelas *Smart Cities* e agrega valor nas construções gráficas no que diz respeito à representação visual dos dados (SOBRAL; GALVãO; BORGES, 2019). Novas estruturas de visualização facilitam a representação geográfica de uma rede urbana e os eventos espaciais que ocorrem nela. Por exemplo, é possível representar graficamente uma rede urbana como assunto principal e usar dados e outras formas visuais possibilitando ao usuário novas perspectivas(SAGL; LOIDL; BEINAT, 2012).

A revisão da literatura apresentado a seguir explorou pesquisas relacionadas a mobilidade urbana em duas categorias, (i) identificação de perfis de regiões e (ii) entender a mobilidade humana. Apesar deste trabalho estar direcionado apenas a identificação de perfis de regiões, explorar pesquisas que têm o objetivo entender a mobilidade humana ajuda a compreender como os pesquisadores tentam entender os objetivos que levam as pessoas se deslocarem em uma região.

3.1 IDENTIFICAR O PERFIL DE REGIÕES EM UMA CIDADE

No trabalho de D'Andrea et al. (2018), os pesquisadores usaram dados de redes sociais e pontos de interesse (POIs) do *Google Maps* (como apresentado na Figura 9), para definir os perfis das áreas da região metropolitana de Milão. Por meio de coleta de dados geolocalizados sobre pontos de interesse, *posts* em redes sociais e *web service*, o trabalho permitiu extrair características significativas para caracterização de áreas da cidades.

Para definir o perfil das áreas da cidade de Milão, os pesquisadores propuseram um *framework*, ilustrado na Figura 10, que consiste em 4 módulos funcionais:

- O módulo *Data Retrieval* é dedicado a coletar dados brutos de diferentes fontes de dados web e armazená-las em um banco de dados.
- O módulo *Data Preparation* prepara o dado para ser analisado em sequência, por exemplo, os dados brutos são filtrados, pre-processados e agrupados, de acordo com a definição da área.
- O módulo *Data Mining Analysis* aplica algoritmo de mineração de dados para o pre-processamento dos dados. O *framework* analisa o *cluster* com o objetivo de encontrar áreas similares na cidade. Os pesquisadores propõe estender o *framework*

[1] https://d3js.org/
[2] https://cloud.google.com/maps-platform/
[3] https://www.mapbox.com/
[4] https://www.carto.com/
[5] https://www.arcgis.com/

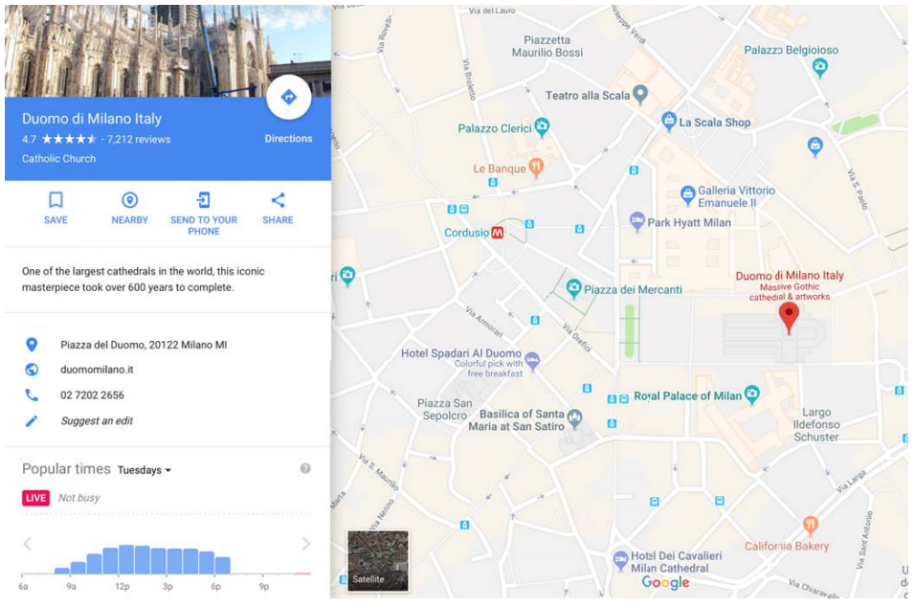

Figura 9 – Uma amostra dos POIs, fornecidos pelo *Google Maps*, na região central de Milão e a esquerda da imagem, mais informações sobre um ponto específico, a catedral de Milão (D'Andrea et al., 2018).

adicionando modelos preditivos, dessa forma, estenderia as funcionalidade do *framework* considerando, por exemplo, detecção de um eventos, estimando dados ocultos e assim por diante.

- Finalmente, o modulo *Result Visualization* oferece uma representação visual e numérica dos dados e do resultado obtido a fim de apoiar a interpretação do resultado.

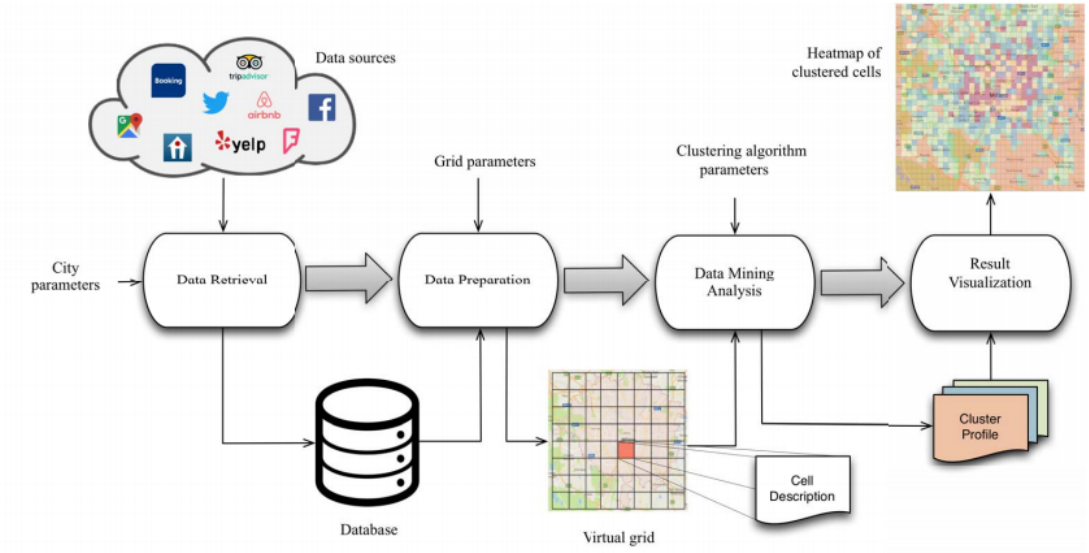

Figura 10 – O *pipeline* do processo da geração de valor através do *framework* proposto com o objetivo de identificar o perfil das diferentes regiões da cidade de Milão (D'Andrea et al., 2018).

Os pesquisadores utilizaram o *framework* na região em torno da cidade metropolitana de Milão, e coletaram aproximadamente 33.680 POIs do *Google Maps*. Foram testadas diferentes partes da cidade, obtido pela configuração das células de diferentes tamanhos que cobrem um determinado quadrado da cidade. O número de células muda de acordo com o tamanho do quadrado, por exemplo, a escolha das células com lados de 500 metros produz uma *grid* com 1188 células, enquanto, células com lados de 300 metros, um total de 3300 células são obtidas. As POIs foram pre-processadas, removendo informação inúteis e mantendo somente informações úteis para a análise e mineração dos dados, o nome, o id, a posição GPS e a categoria correspondente. Cada POI é associado com a célula correspondente, de acordo com a posição no GPS. Assim, cada característica é extraída e cada célula é descrita de acordo com o número de POIs associado as células.

Com intuito de simplificar o processo e para uma melhor interpretação dos resultados, no módulo *Data Preparation* os pesquisadores agruparam as categorias em 4 macro-categorias como mostrado na Tabela 2. A categoria *DomCat* representa a categoria predominante no *cluster*, ou seja, aquela que mais se repete em todo o conjunto. Já a categoria *DistCat* diz respeito a categoria que se repete com mais frequência em cada célula do *cluster*, não necessariamente à que mais ocorre e sim a que está mais presente em cada célula do *cluster*. A categoria *POIDen* diz respeito à quantidade média de POIs em cada célula, quanto mais POIs, mais densa é a célula. Por fim, a categoria *ClNum* representa à quantidade de células por *cluster*.

Cada célula foi rotulada de acordo com a quantidade e similariedade dos POIs, para isso os pesquisadores usaram o algorítimo *K-means* (variando o valor de k de 3 a 7), para identificar *clusters* com POIs similares. Os pesquisadores usaram k = 4, pois segundo eles, este valor produziu facilmente a interpretação do *cluster*. O resultado é resumido na Tabela 2, e nas Figuras 11 e 12. Além do k *clusters* gerados pelo algoritmo *K-means*, os pesquisadores também adicionaram um *cluster* vazio, ou seja, *cluster* que não há POIs em sua área, o que indica uma área residencial de Milão. Portanto, o número total de *clusters* é igual a 5. Além disso, os pesquisadores entenderam que o melhor particionamento da *grid* em termo de interpretabilidade resultou ser aquele que tem um tamanho dos lados da célula em 500 metros, conforme mostrado na Figura 11. De fato, as células menores aumentam a complexidade e apresenta uma excessiva granularidade, enquanto que, células maiores tornam maior a diferença entre as áreas.

Os pesquisadores chegaram a conclusão que o *framework* desenvolvido apresenta-se como uma boa técnica com intuito de definir perfis de uma determinada cidade, porém o trabalho se limitou a estudar o perfil das regiões da cidade de Milão e não ofereceu indícios que é possível analisar outras grandes metrópoles. Expandir tal ferramenta para analisar perfil de regiões em outras cidades pode ser útil em pesquisas para entender o deslocamento das pessoas. Conforme será tratado na Seção 3.2, pesquisas tem se esforçado

Interpretação das Medidas				
Cluster ID (cor)	**DomCat**	**DistCat**	**POIDen**	**ClNum**
#1 (azul escuro)	-	-	0	160
#2 (azul claro)	Lojas, Serviços	Saúde, Transportes	41.7	364
#3 (verde)	Serviços, Lojas	Serviços, Finanças	51	228
#4 (laranja)	Serviços, Transportes	Transporte, Entretenimento	13.9	420
#5 (vermelho)	Lojas, Restaurantes	Lojas, Restaurantes	57.6	16

Tabela 2 – Para melhor interpretação da representatividade de cada *cluster*, 5 macro-categorias foram criadas e suas medidas servem para identificar o perfil de cada região de Milão (D'Andrea et al., 2018).

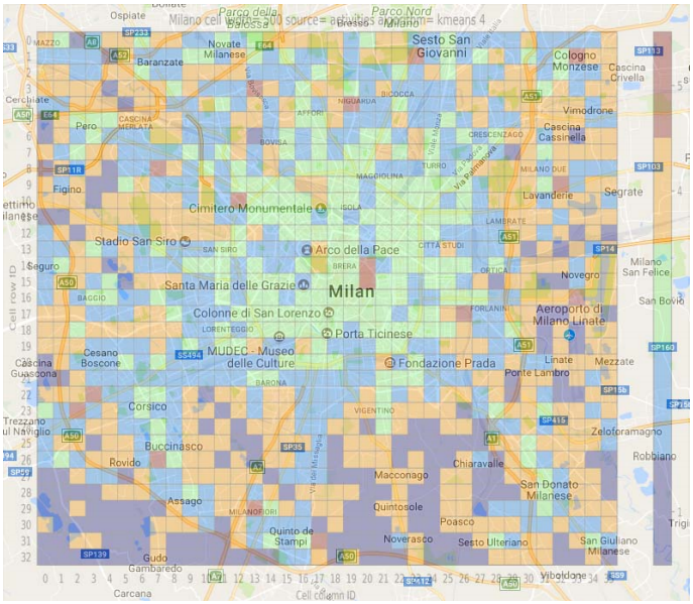

Figura 11 – Resultado do agrupamento em um *heatmap* com 5 *clusters*, sobre o mapa da cidade de Milão (Fonte: *Google Maps*) com 1188 células com tamanho de 500 metros cada lado (D'Andrea et al., 2018).

para traçar a localização dos usuários baseado em registro de chamadas de celulares (CDR), mas enquanto esses dados de CDR informam uma precisão grosseira sobre localização, dados de redes sociais baseados na localização fornecem uma localização mais precisa, por exemplo é possível distinguir se um *check-in* foi realizado no 1º ou no 2º andar de um prédio (CHO; MYERS; LESKOVEC, 2011).

Para apoiar o estudo da mobilidade humana, diversos pesquisadores tem buscado entender qual a relação entre as redes sociais baseada em localização (LBSN) e o deslocamento humano, no trabalho de Backstrom, Sun e Marlow (2010), foi possível observar e medir diretamente a relação entre geografia e amizade a fim de prever a localização de um

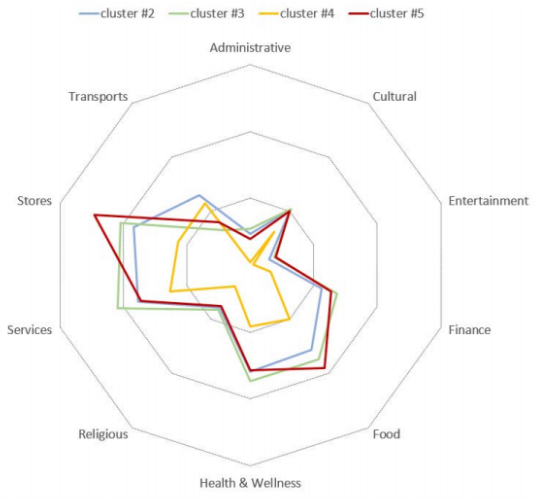

Figura 12 – Um gráfico de radar com 4 *clusters* em escala logarítmica para a macro-categoria *DomCat*(D'Andrea et al., 2018).

indivíduo a partir de um range de IP's em uma determinada área geográfica.

Já, outros trabalhos têm procurado entender a partir dos dados, qual é a atividade que o usuário está realizando baseado em sua posição geográfica e para qual objetivo ele se desloca. Fazer tais descobertas ainda é um dos grandes desafios para análise da mobilidade urbana. No trabalho de Xiong et al. (2014), os pesquisadores utilizaram dados produzidos por telefonia móvel (CDR), para entender tais deslocamentos.

Diferentemente de dados de redes sociais ou pesquisas com grupos de usuários, que possuem alguma forma de rótulo com informações se o indivíduo está em casa, no trabalho, em alguma atividade de lazer, em compras, ou em outras atividades, os dados de CDR não possuem informações se a posição geográfica do usuário representa uma atividade específica. Para rotular a localização do usuário os pesquisadores utilizaram o algoritmo proposto nos trabalhos de Hung e Peng (2011) e Wang et al. (2010) para definir uma parada, tendo em vista que dados de CDR brutos podem possuir vários registros de um único local e que esses dados não são precisos. É possível visualizar na Figura 13 que existe vários registros muito próximos em dois momentos no gráfico, o que identifica que o usuário estava em um mesmo local no primeiro momento em um intervalo de tempo de até 3, houve um deslocamento até um segundo momento que iniciou próximo ao intervalo de tempo 3.5 e seguiu até 5.

Com o intuito de melhorar a acurácia na identificação de locais de atividades afim de entender o comportamento humano, alguns pesquisadores tem concentrado esforços para combinar dados de CDRs com dados geolocalizados produzidos em aplicativos *web* como redes sociais. Dados de CDRs possuem baixa precisão na localização do usuário, pois a localização é estimada de acordo com a latência do sinal da torre de transmissão, diferentemente dos dados produzidos por sistemas de geolocalização como GPS que

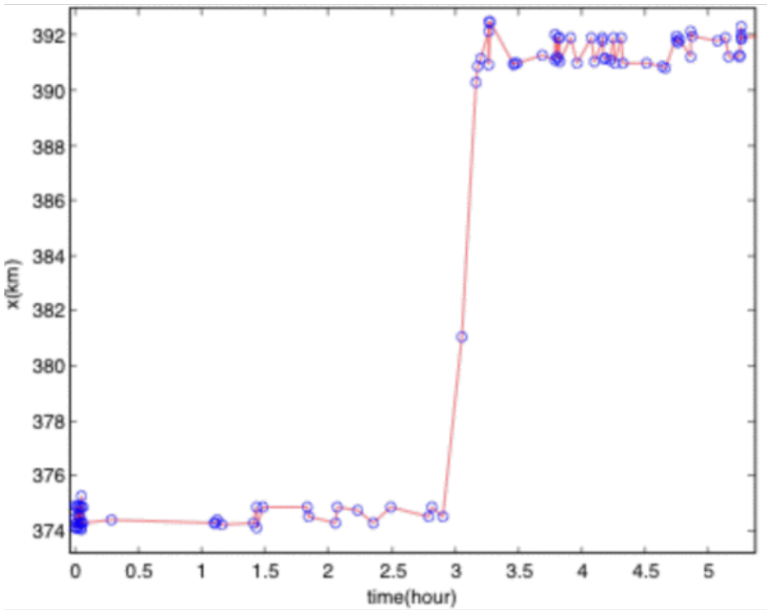

Figura 13 – Pontos de registro da posição de um usuário em cada observação dos dados de CDR. É possível identificar 2 grupos de registros, isso demonstra que em cada evento de telefonia móvel em um mesmo local, foram geradas posições geográficas não tão precisas (WANG et al., 2010).

possuem uma maior precisão na localização. A grande vantagem do uso de dados de CDRs é quantidade de dados gerada. As grandes cidades possuem uma maior cobertura das torres de transmissão de telefonia móvel e a grande maioria dos moradores possuem um telefone móvel. Já dados de GPS é necessário que o usuário possua um *smartphone* e o serviço de GPS esteja ativado.

Também, com o mesmo objetivo de combinar dados de GPS e de CDRs, no trabalho de Noulas, Mascolo e Frias-Martinez (2013), os pesquisadores exploraram padrões nas proximidade das torres de transmissão e combinaram com dados de milhões de locais do aplicativo *Foursquare*, produzidos usuários com a localização de GPS, para inferir o tipo de atividade nos bairros de Barcelona na Espanha, a fim de entender as atividades humanas. Além disso, os pesquisadores analisaram os padrões de *check-in* de usuários do *Foursquare* em locais geograficamente próximos às torres e exploraram as anotações semânticas dos *check-ins* para caracterizar as áreas próximas em termos de atividades urbanas conhecidas, como alimentação, vida noturna, trabalho e viagens, entre outras. Ainda, os pesquisadores formularam uma tarefa de aprendizado supervisionado que visa criar uma função de aprendizado para associar o sinal de telecomunicação de entrada de uma torre às categorias do *Foursquare* de locais mais próximos. Segundo os pesquisadores, isso potencialmente levaria à acurácia para caracterização do perfil da áreas geográficas.

Além de extrair características necessárias para rotular o sinal de telecomunicações de uma torre, foi realizada a etapa de pré-processamento de agregação de torres geograficamente próximas a super-torres virtuais. O objetivo desta consiste em: primeiro e

mais importante, em ambientes urbanos muito densos, existe um número muito grande de torres de telecomunicações que separam o espaço geográfico em partições muito pequenas. Às vezes, duas ou três torres podem estar quase presas umas às outras para servir uma área, como é possível ver na Figura 14. O efeito disso é que, para cada torre, há poucos locais associados ao *Foursquare*, assim, a extração de rótulos de uma amostra tão pequena resulta em ruído. Sendo assim, por meio da agregação, é combinado o sinal de torres próximas, desta forma melhorando a qualidade do sinal de entrada, essencial para a tarefa de aprendizado.

Figura 14 – Ilustrando o efeito da agregação de torres em super torres que cobrem áreas geográficas maiores. As coordenadas geográficas das torres foram agrupadas para produzir centróides espaciais que cobrem áreas maiores. Cada célula de voronoi é colorida com a atividade mais popular, explorando a popularidade de locais do *Foursquare* próximos (Noulas; Mascolo; Frias-Martinez, 2013).

Para realizar a tarefa agregação da torres, os pesqusiadores aplicaram o algoritmo de agrupamento espacial baseado em densidade chamado DBSCAN (ESTER et al., 1996). O algoritmo recebe como entrada coordenadas de latitude e longitude das torres e retorna os agrupamentos de torres geograficamente próximas. Foi definido o centróide de cada *cluster* para ser uma super-torre, isto é, uma torre cujo sinal de entrada é definido por agregar os registros das torres que foram agrupadas pelo algoritmo. Na Figura 14, é mostrado o efeito da agregação de torres na cidade de Madri. À esquerda, é possível observar que o mosaico de Voronoi ilustrando o conjunto original de torres na cidade sem considerar nenhum mecanismo de agregação. No lado direito da figura, é mostrado o mosaico após a aplicação da etapa de agregação. Existem duas observações principais. Primeiro e por definição, as áreas periféricas, após a agregação são maiores. Segundo, a distribuição de atividades populares em áreas da cidade muda (veja cores diferentes mapeando as atividades urbanas).

3.2 ANÁLISE DE PADRÕES DE ATIVIDADE HUMANAS

Analisar padrões de atividades humanas tem originado diversos trabalhos para planejamento urbano a fim de melhor a logística de transporte para escola, trabalho, lazer, etc. Um exemplo é o trabalho de Jiang, Ferreira e González (2012) que desenvolveram uma pesquisa sobre trajetos dos moradores da área metropolitana de Chicago, foram analisados mais de 30.000 indivíduos que participaram de uma pesquisa que durou de 1 a 2 dias e foi realizado no período de janeiro de 2007 a fevereiro de 2008.

Para registrar o local onde o indivíduo está, o dia foi dividido em 288 partes de 5 minutos e cada uma das 288 partes possuía a identificação da atividade e as coordenadas (latitude e longitude) de onde o indivíduo se encontra. Vale ressaltar que, se em um intervalo de 5 minutos, o indivíduo identificasse duas ou mais atividades, somente a primeira atividade registrada seria considerada.

Para fins de análise, foram consideradas 22 principais atividades dos dados originais da pesquisa, as demais entraram em uma categoria denominada "Outros", totalizando assim 23 atividades. De acordo com o padrão proposto por Bowman e Ben-Akiva (2001) para estudos urbanos, as 23 atividades foram agrupadas em menos tipos de atividades e identificadas por uma cor específica para cada atividade conforme mostrado na Figura 15.

Aggregated Activity Types	Original Primary Trip Purposes
Home	1. Working at home (for pay); 2. All other home activities
Work	3. Work/Job; 4. All other activities at work; 11. Work/Business related
School	5. Attending class; 6. All other activities at school
Transportation Transitions	7. Change type of transportation/transfer; 8. Dropped off passenger from car; 9. Picked up passenger; 10. Other, specify-transportation; 12. Service private vehicle; 24. Loop trip
Shopping/Errands	13. Routine shopping; 14. Shopping for major purchases; 15. household errands
Personal Business	16. Personal Business; 18. Health Care
Recreation/Entertainment	17. Eat meal outside of home; 20. Recreation/Entertainment; 21. Visit friends/Relatives
Civic/Religious	19. Civic/Religious activities
Other	97. Other

Figura 15 – Padrão proposto por Bowman e Ben-Akiva (2001) relacionando as 23 atividades auto referidas em 9 grupos de atividades aplicadas no trabalho de (JIANG; FERREIRA; GONZÁLEZ, 2012).

A partir do agrupamento das atividades dos indivíduos, os pesquisadores procuraram responder a três questões: a) Qual a estrutura das diárias inerente aos indivíduos em uma área metropolitana?, e b) Variação das atividades diárias individuais?, e c) Agrupamento de comportamentos individuais baseado na similaridade de suas atividades diárias?

Para responder essas perguntas, os pesquisadores utilizaram dois métodos *Principal Component Analysis* (PCA) e o algoritmo de agrupamento *K-means*. O PCA é usado nesse trabalho para obter a decomposição dos valores da pesquisa em uma matriz de covariância. Um exemplo do uso do PCA para decompor valores em uma matriz de covariância foi usado no trabalho de Hastie et al. (2005). Para cada uma das amostras individuais da pesquisa, foi calculado o desvio padrão e a sua variância para posteriormente gerar a matriz de covariância para saber o quão longe ou próximo da média de população está cada individuo da amostra. Dessa forma, os pesquisadores conseguiram responder as questões "a" e "b". Para a questão "c", foi usado o algoritmos de agrupamento *K-means*, com intuito de particionar grupos de indivíduos com atividades similares. No conjunto de variáveis para aplicação do algoritmo *K-means*, as variáveis com registro de data e hora não representam precisamente a data e hora do evento, e sim a data e hora do início do intervalo de 5 minutos para cada um dos 288 intervalos de um único dia. Ainda, neste trabalho os pesquisadores desenvolveram uma ferramenta de animação temporal para visualização do deslocamento e atividades (diferenciadas por cores, conforme o padrão proposto por Bowman e Ben-Akiva (2001) dos indivíduos pesquisados em um dia útil da semana. A Figura 16 mostra 4 *snapshots* da ferramenta de animação temporal em diferentes horários do dia 06h00, 12h00, 18h00, 00h00.

Figura 16 – *Snapshots* obtidos da ferramenta de animação temporal demonstrando atividades humanas em diferentes horas do dia em um dia da semana em Chicago (JIANG; FERREIRA; GONZÁLEZ, 2012).

Apesar dos resultados deste trabalho comprovarem a alta previsibilidade da mobilidade humana estudando dados públicos sobre trajetos na área metropolitana de Chicago e com o uso do método PCA e o algoritmo de agrupamento *K-means*, que são popularmente reconhecidos como técnicas eficazes na análise da correlação entre as variáveis e na identi-

ficação de grupos similares respectivamente – os 30.000 registros estudados correspondem a uma pequena fatia da população de Chicago e a técnica apresentada pelos pesquisadores poderia resultar em dados mais confiáveis se usada em volumes de dados maiores.

Em seu trabalho mais atual sobre padrões de mobilidade urbana, Jiang, Ferreira e Gonzalez (2017) coletaram durante 14 dias consecutivos os dados de telefonia móvel (CDRs), na cidade-estado de Cingapura para examinar os padrões de mobilidade de indivíduos anônimos na região metropolitana. O conjunto de dados do período estudado contém 3,17 milhões de usuários de celulares anônimos e um total de 722,92 milhões de registros de uso de telefones. Existem mais de 5 mil torres de celular em Cingapura, com um espaçamento de cerca de 50 metros no denso centro da cidade a alguns quilômetros na região suburbana. Apesar da alta penetração os dados de CDRs foram combinados com dados do censo da cidade de Cingapura que inclui a população de diferentes grupos demográficos no nível da zona de planejamento.

Para entender os padrões de mobilidade humana no nível metropolitano para fins de planejamento urbano e de transporte Jiang, Ferreira e Gonzalez (2017) desenvolveu um pipeline (veja na Figura 17) para (1) análise de dados CDR e extrair locais de permanência de usuários de telefone; (2) detectar a localização residencial dos usuários de telefone; (3) filtrar usuários e selecionar amostras estatisticamente representativas dos dados analisados do CDR; (4) identificar as redes de mobilidade diárias dos usuários de telefone; (5) derivação de fatores de expansão para as observações filtradas do telefone, combinando dados processados do telefone e os dados do censo; expandir usuários de celular, viagens e motivos diários; e agregando-os da torre às zonas de análise de transporte.

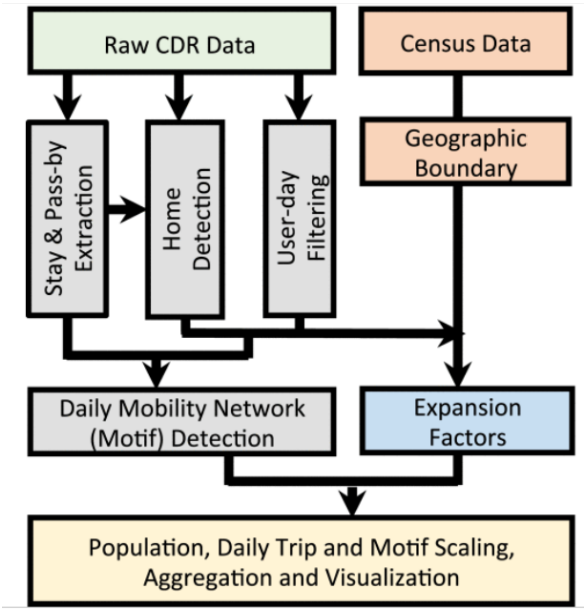

Figura 17 – *Pipeline* do processo de descoberta de conhecimento combinando dados de CDR com censitários de Cingapura (Jiang; Ferreira; Gonzalez, 2017).

A partir dos dados coletados e combinados, foi possível desenvolver uma série de

técnicas de visualização para descobrir padrões humanos. Por exemplo, na Figura 18, é apresentada a distribuição espacial dos fatores de expansão do usuário no nível da torre em Cingapura, ilustrados por diferentes cores, com base no alcance da medida do desvio padrão. É possível ver que os fatores de expansão do usuário são mais altos na parte norte de Cingapura, o que significa que a proporção de usuários de telefone e população nessa região é menor que a média da cidade.

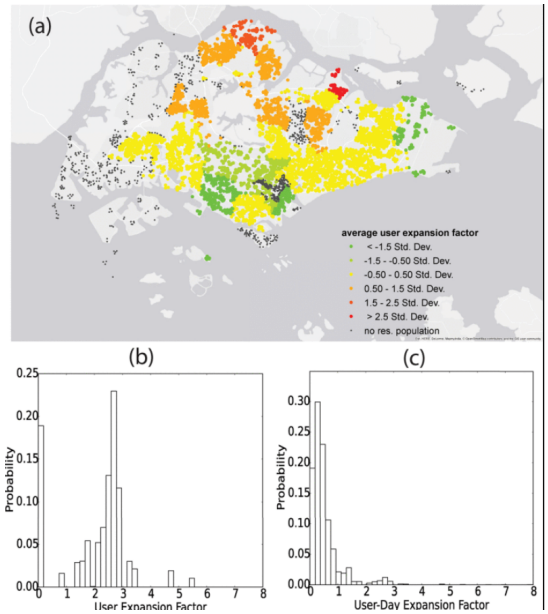

Figura 18 – (a) Distribuição espacial dos fatores de expansão do usuário no nível da torre e distribuição de frequência (b) fatores de expansão do usuário e (c) fatores de expansão no dia do usuário. (Jiang; Ferreira; Gonzalez, 2017)

Outros pesquisadores tem proposto métodos de análise da mobilidade urbana a partir de dados coletados de redes sociais, como é o caso do trabalho de Jurdak et al. (2015), os pesquisadores proporam o uso do *Twitter* como um *proxy* para a mobilidade humana, pois ele se baseia em dados publicamente disponíveis e fornece posicionamento de alta resolução quando os usuários optam por marcar geograficamente seus *tweets* com sua localização atual.

Foram analisados um grande conjunto de dados com 7.811.004 *tweets* de 156.607 usuários do *Twitter* de setembro de 2013 a abril de 2014 na Austrália para determinar quão representativos são os padrões de mobilidade baseados no *Twitter* de movimentos populacionais e de movimento individual. Os pesquisadores compararam os padrões de mobilidade observados no *Twitter* com os padrões observados em outras tecnologias, como registros de dados de chamadas. Para essa análise foi utiliza indicadores universais para caracterizar padrões de mobilidade de *tweets* com identificação geográfica, ou seja, a distribuição de deslocamento e a distribuição do raio de rotação que mede a distância que os indivíduos normalmente se deslocam (sua órbita espacial).

Os pesquisadores concluiram que os *tweets* com informações geográficas podem

capturar características ricas da mobilidade humana, como a diversidade de movimento de indivíduos dentro e entre cidades. Também concluiram que os motoristas de curta distância passam a maior parte do tempo em grandes áreas metropolitanas, em contraste com os movimentos de motoristas de distância intermediária, refletindo o impacto de diferentes modos de viagem. A partir das análise deste estudo, há evidências sólidas de que o *Twitter* pode realmente ser um proxy útil para rastrear e prever o movimento humano.

3.3 IDENTIFICAÇÃO DE FLUXOS DO TRÁFEGO DE VEÍCULOS

A análise e monitoramento de fluxos de tráfego de veículos é um tópico muito estudado em visualização de mobilidade urbana (SOBRAL; GALVãO; BORGES, 2019). No trabalho de Andrienko et al. (2013) foi proposto uma aplicação de diagramas de ruas na análise espaço-temporal para análise do congestionamento de tráfego urbano (veja na Figura 19). Como em uma flor, o diagrama de rosas (do inglês *rose chart*), os segmentos do círculo representam as horas do dia. E tais segmentos são usados para representar o número de congestionamentos de tráfego e o tamanho do segmento representa a duração em tempo dos engarrafamentos. A técnica de visualização foi aplicada com dados coletados do transporte público da cidade de Helsinque na Finlândia.

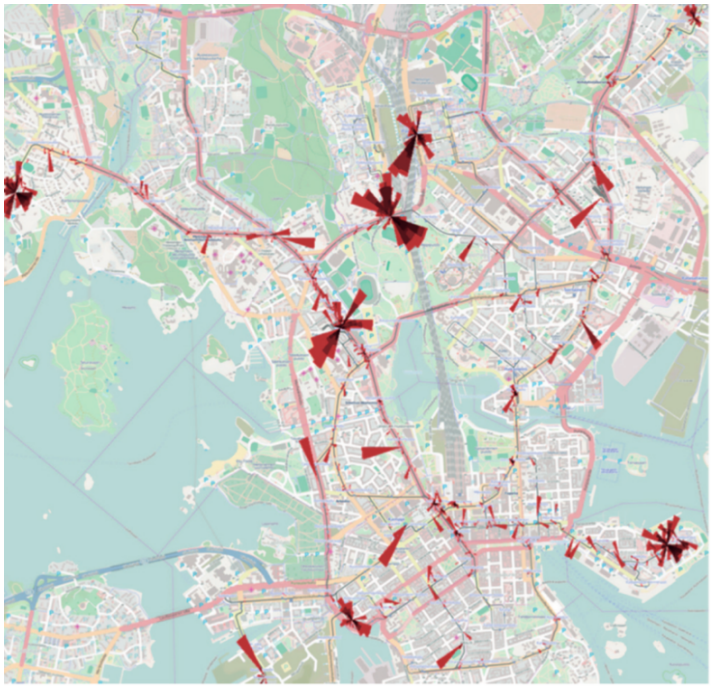

Figura 19 – Diagrama de rosas aplicado a entender o congestionamento do tráfego na cidade de Helsinque. Cada diagrama de rosas foi definido por um usuário especialista no trânsito de Helsinque em ponto estratégicos da cidade (ANDRIENKO et al., 2013).

Uma técnica de visualização vista com frequência na literatura são os *heatmaps* ou mapas de calor, os autores utilizam os mapas de calor para identificar padrões anormais no monitoramento do trânsito. Um exemplo de trabalho que usou mapa de calor para analisar o congestionamento foi o de Song e Miller (2012), que utilizou uma matriz de mapa de calor para analisar os padrões de congestionamento em duas granularidades temporais: dias das semanas ou meses e hora do dia, conforme mostrado na Figura 20. Tais matrizes podem ser eficazes na identificação de padrões anormais e têm sido aplicadas a outras ferramentas de visualização. Liu et al. (2013) e Pu et al. (2013), aplicaram mapas de calor circulares sobreposto em um mapa.

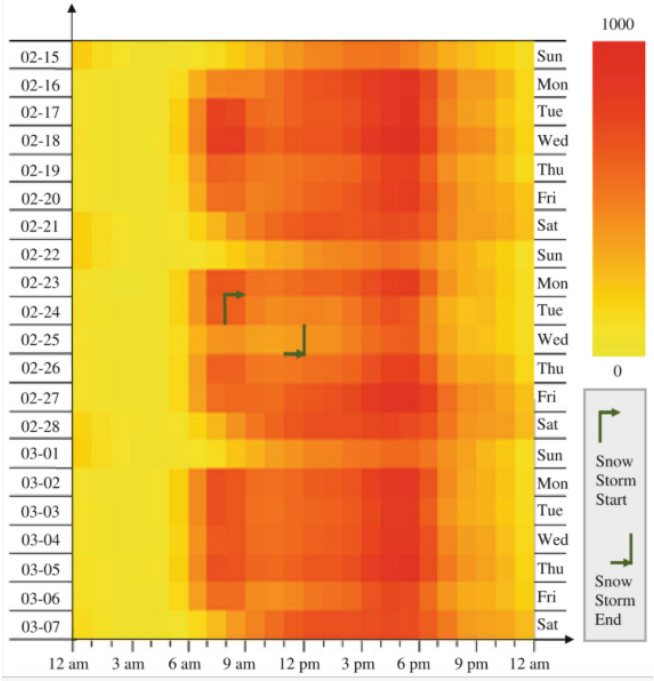

Figura 20 – Visualização da matriz do mapa de calor para análise de congestionamento de tráfego (SONG; MILLER, 2012).

Mapas de calor podem ser combinados com um mapa geográfico, apresentado ao usuário a real percepção do local de maior incidência de congestionamento. Nos exemplos da Figura 21, a imagem (a) consiste na sobreposição de um mapa de calor, equanto (b) fornece cores aos segmentos da estrada de acordo com uma determianda escala (POCO et al., 2015).

O trabalho de Chen et al. (2015b) destaca a importância da interação do zoom semântico para a análise de fenômenos em diferentes níveis. Os gargalos de velocidade recuperados dos sensores do veículo podem ser analisados em várias granularidades temporais em uma matriz de mapa de calor. Conforme o usuário seleciona o período de tempo desejado, a técnica de visualização se adapta para mostrar as respectivas velocidades do veículo e intensidade do fluxo.

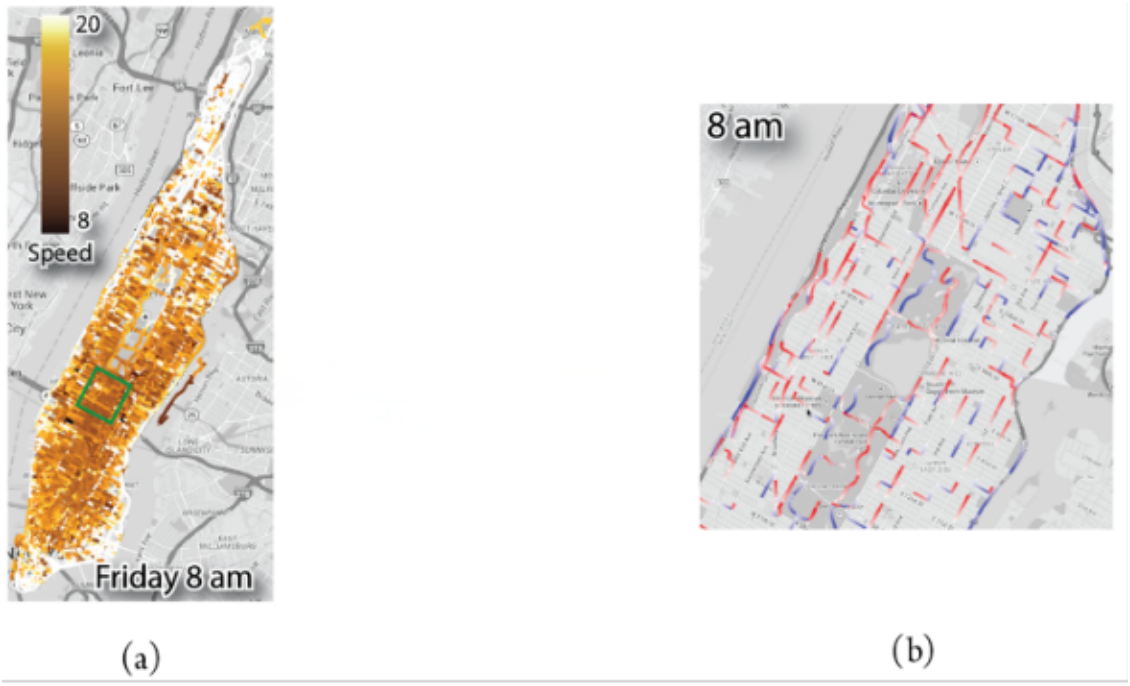

Figura 21 – Outras formas de representação do mapa de calor sobreposto em um mapa geográfico (POCO et al., 2015).

3.4 *ANÁLISE DA DINÂMICA DO DESLOCAMENTO DE PESSOAS*

Estudar a dinâmica que leva pessoas a irem de um ponto A a um ponto B têm se concentrado em pontos críticos urbanos. No trabalho de Sagl, Loidl e Beinat (2012), os pesquisadores procuraram compreender melhor os padrões espaço-temporais típicos da mobilidade humana coletiva na escala operacional de uma cidade e sua periferia próxima, a pesquisa analisou quatro cidades italianas, Trieste, Udine, Pordenone e Gorizia, assim o trabalho foi capaz de revelar semelhanças e diferenças na configuração funcional das cidades em termos de mobilidade. Nesse trabalho os pesquisadores criaram uma matriz visual para analisar a correlação entre as quatro cidades em diferentes dias da semana, as linhas coloridas identificam o tipo de atividade em diferntes horas do dia (veja na Figura 22).

Demissie, Correia e Bento (2013) analisaram dados produzidos por telefonia móvel, especificamente, dados conhecidos como *call details record* (CDRs), para entender o processo de transferência de uma chamada em movimento de aparelhos de telefonia móvel, ou seja, quando uma ligação ativa é comutada de uma torre de transmissão para outra. O objetivo foi destacar pontos críticos na cobertura do sinal da torre, detectar pontos de congestionamento de celulares e padrões de mobilidade humana. A exploração visual foi aplicada com o uso de mapas de fluxo, como apresentado na Figura 23, na qual as setas representam as direções da transferência de uma torre para outra, a densidade das setas

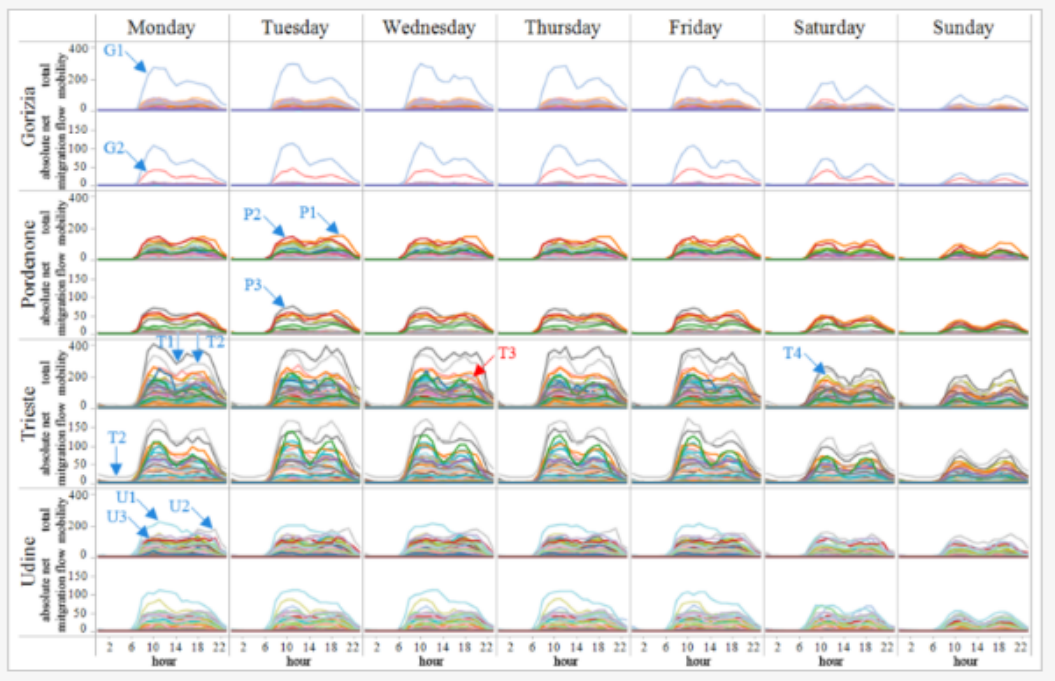

Figura 22 – Visualização apresenta a correlação de atividades urbanas entre cidades italianas em diferentes dias da semana (SAGL; LOIDL; BEINAT, 2012).

representam o volume de transferencia em um período e o triângulo em verde representa a torre de transmissão. Já na Figura 24, os círculos dimensionados representam se a torre passou a enviar o sinal de telefonia (entrega de entrada) para um dispositivo em uma chamada ativa ou se houve o inverso, isto é, se a torre interrompeu o envio do sinal pois o dispositivo já não está mais na área de cobertura (entrega de saída).

No trabalho de Chen et al. (2015a) foi desenvolvido uma técnica para análise visual interativa com o ojetivo de analisar padrões de movimento de usuários de mídia social. O processo de análise visual é baseado em um modelo de incerteza que os pesquisadores desenvolveram para estimar os atributos temporais dos dados de mídia social. Por meio desse procedimento, os usuários podem extrair informações mais confiáveis sobre o tempo de viagem dos dados e criar categorias razoáveis de movimento para explorar ainda mais os padrões de movimento. O sistema apresenta várias técnicas abstratas de visualização, conectadas a interações como selecionar sub conjuntos dos dados e vinculação entre os gráficos, que são combinadas com uma visualização baseada em mapa para exibir dados espaciais e temporais agregados.

3.5 *ANÁLISE DE INCIDENTES DE TRÁFEGO*

De acordo com Albino, Berardi e Dangelico (2015), é comum nas grande cidades o registro de incidentes com informações detalhadas como a hora e a localização, número de veículos envolvidos, fechamento de faixas, condições meteorológicas e das estradas,

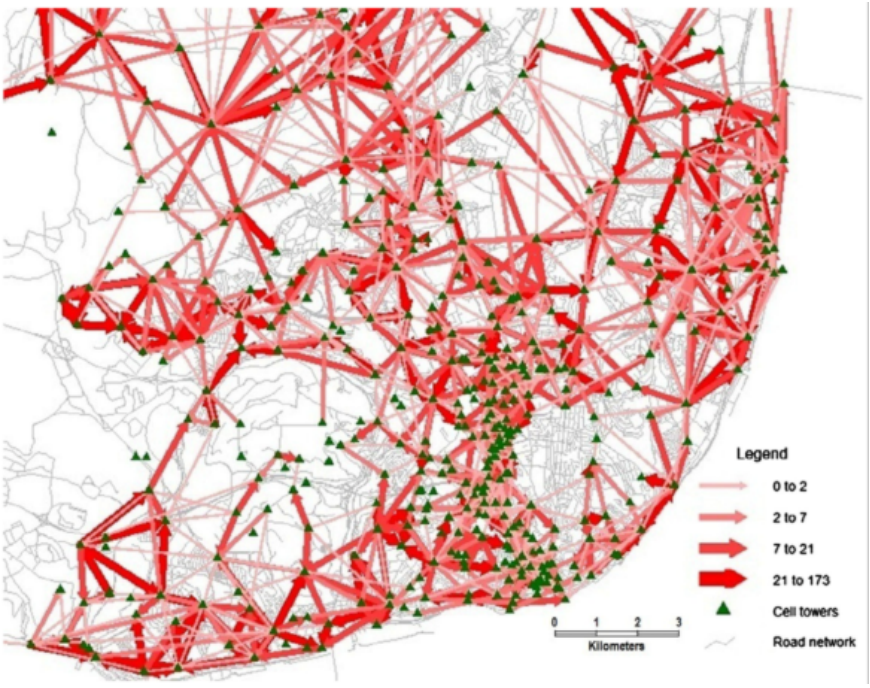

Figura 23 – Visualização baseada em mapas de fluxo de transferência de torres de telefonia móvel durante uma chamada telefônica ativa (DEMISSIE; CORREIA; BENTO, 2013).

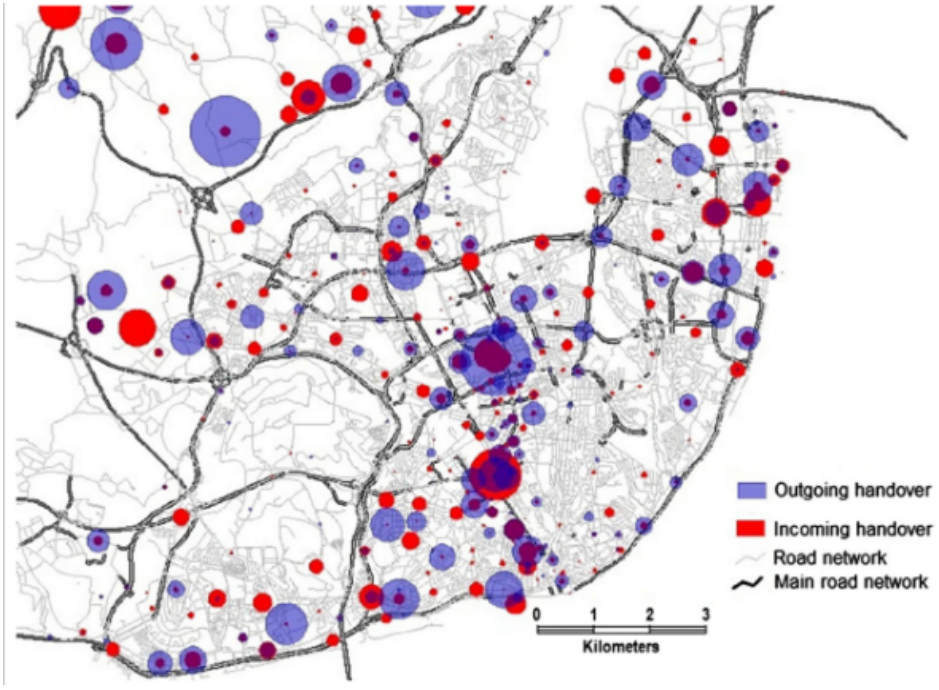

Figura 24 – Visualização baseada em mapa de volume de entrega de entrada e saída usando círculos dimensionados (DEMISSIE; CORREIA; BENTO, 2013).

gravidade do incidente, etc. Embora exista um consenso dos orgãos responsáveis pela administração das vias públicas para desenvolvimento de padrões para armazenar dados de incidentes, pouco esforço foi feito para projetar ferramentas de análise visual apropriadas para analisar os dados, extrair conhecimento significativo e representar resultados.

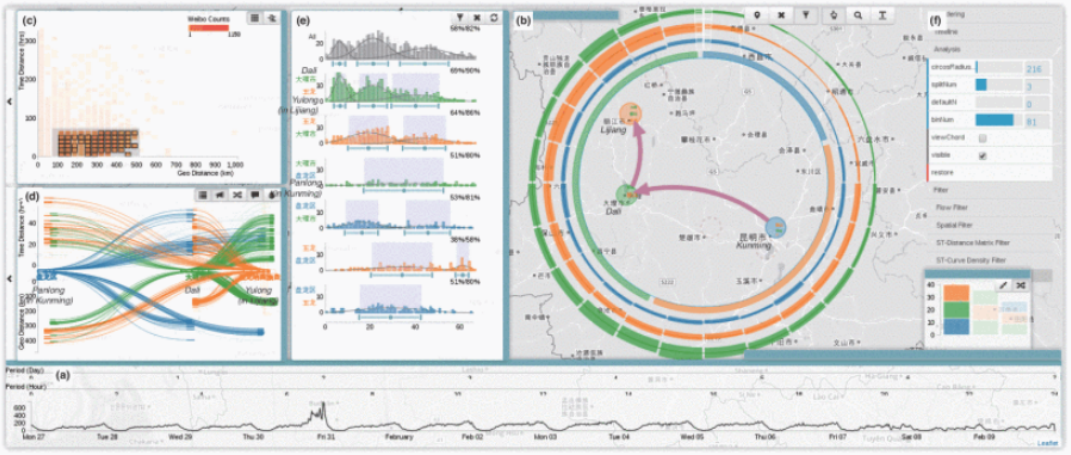

Figura 25 – Um sistema de análise visual proposto por Chen et al. (2015a) para análise visual de dados esparsos de *microblogging*. Várias técnicas de visualização estão inter-relacionadas. Gráficos de tempo (a) e matrizes de mapa de calor (c) representam a distribuição do movimento na distância e no tempo. A visualização baseada em mapas em (b) exibe fluxos de espaço e tempo agregados entre cidades. Finalmente, (d) apresenta o diagramas de Sankey para representar movimento pareados no tempo.

No trabalho de Pack et al. (2009), foi desenvolvido uma ferramenta de análise visual, baseada na Web, chamada ICE (*Incident Cluster Explorer*), é proposta como um aplicativo que oferece uma análise sofisticada e fácil de usar dos conjuntos de dados de incidentes de transporte. A ferramenta fornece ao usuário um conjunto intuitivo de funcionalidades que inclui filtragem de dados, visualizações geoespaciais, funções de classificação estatística e recursos multidimensionais de exploração de dados.

Na Figura 26, é apresentada a ferramenta ICE que faz a interligação entre as áreas de visualização. No mapa, cada círculo representa um incidente em uma determinada região, o usuário pode clicar em uma barra no histograma na parte inferior da ferramenta que os incidentes relacionados a essa categoria de incidente são destacados em larajando na interface do mapa. Além disso, o usuário também pode selecionar um conjunto de incidentes no mapa e os incidentes selecionados são destacados no histograma como um subconjunto destacado.

Pode-se também examinar as frequências de incidentes de maneira mensal, semanal, diária ou horária. Primeiro, clicar em certos tipos de histogramas destaca os incidentes associados a esses valores no mapa. Além disso, os dados temporais são representados como uma hierarquia de valores (minutos, horas, dias, meses, anos), e os histogramas temporais no ICE suportam uma maneira lógica de interagir com essa hierarquia. A visualização inicial exibe a correlação entre os números de incidentes e o mês do ano em que esses incidentes ocorreram.

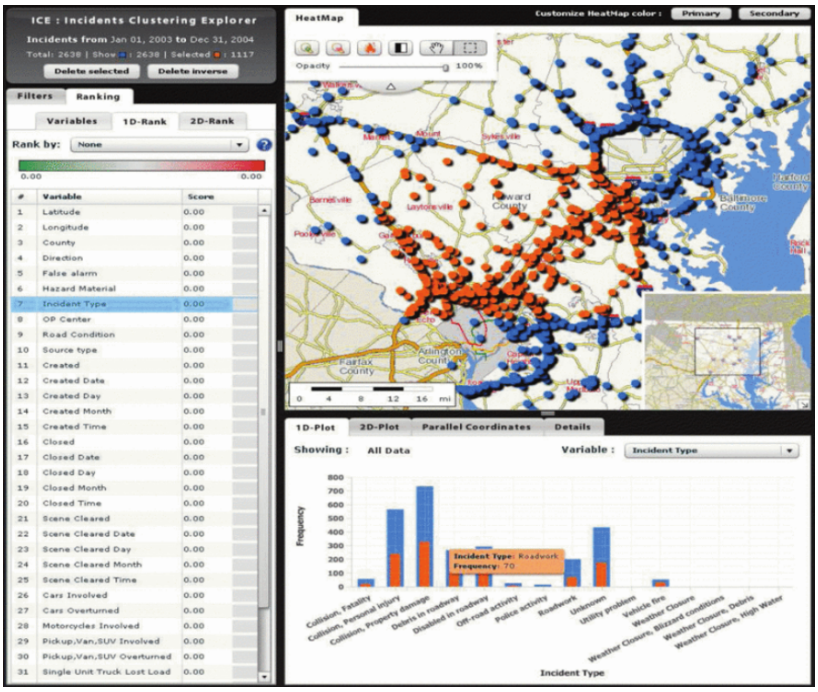

Figura 26 – Ferramenta ICE mostrando os histogramas interativos e a interligação entre os mapas (PACK et al., 2009).

4 PLACEPROFILE: DESCOBRINDO PADRÕES BASEADOS EM PONTOS DE INTERESSE

Este capítulo apresenta uma aplicação web denominada PlaceProfile, a qual usa informação de pontos de interesse para criar perfis e rotular áreas de uma cidade. Para tanto, o PlaceProfile emprega metáforas visuais e algoritmos de agrupamento para auxiliar usuários a identificar as principais atividades de diferentes regiões de uma cidade ou região metropolitana. Na Figura 27 mostra a arquitetura do PlaceProfile, que consiste em quatro componentes funcionais: Preparação; Coleta de dados; Mineração de Dados; e Visualização.

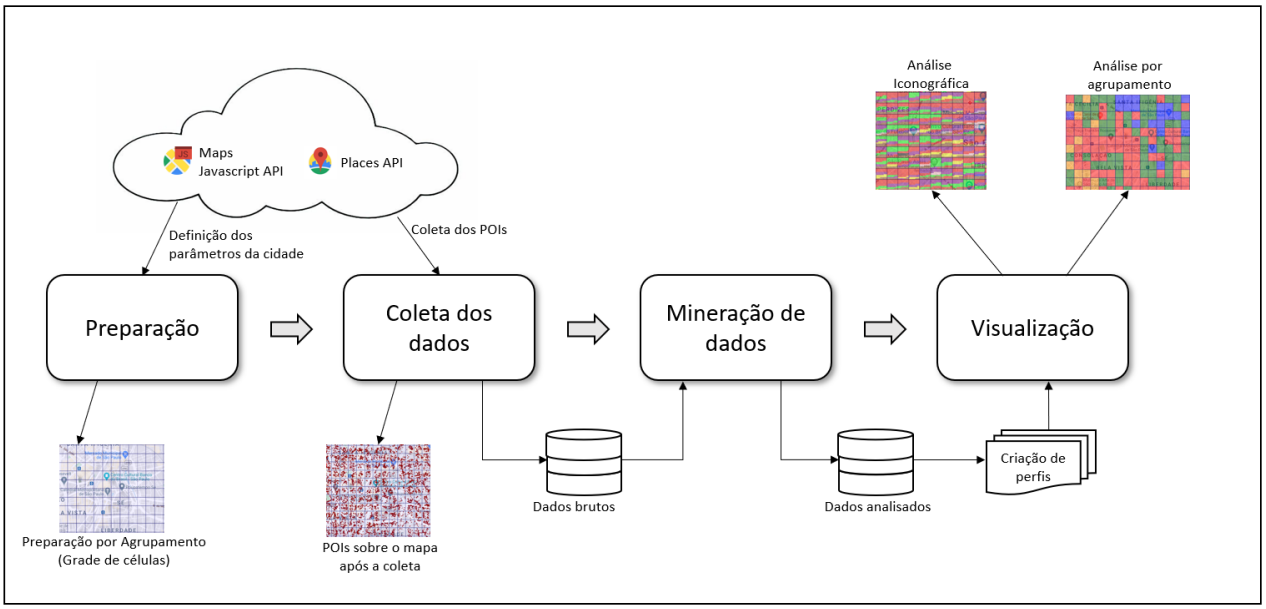

Figura 27 – PlaceProfile consiste em uma aplicação web para mineração visual de dados após a coleta dos pontos de interesse no Google Maps. Os principais componentes do PlaceProfile são preparação, coleta de dados, mineração de dados e a visualização.

4.1 PREPARAÇÃO

A etapa de preparação consiste em definir parâmetros que serão usados para a etapa coleta dos dados. Primeiramente o usuário deve definir a região que ele deseja analisar, com o uso do recurso de zoom o usuário pode controlar com mais precisão a área de cobertura desta região, a qual é dividida em células, ou seja, áreas quadriculares de um mesmo tamanho para formar uma grid, a qual é sobreposto à região de interesse. Cada célula da grid será identificada como uma micro região, com largura e altura em metros de acordo como definido pelo usuário. Obviamente que, quanto maior o tamanho da célula, menor será a quantidade de células na grid e quando maior for a região que se deseja analisar, maior será a quantidade de células na grid. Na Figura 35, é mostrado a grid plotada sobre o mapa de uma área da cidade de São Paulo. Nesse exemplo, foi definido o tamanho de 250x250 metros para cada célula.

Figura 28 – PlaceProfile: Neste exemplo, uma grid foi plotada sobre o mapa de uma área da cidade de São Paulo, cada célula da grid tem um tamanho de 250x250 metros, totalizando 476 células.

A biblioteca Maps Javascript API do Google foi muito útil ao definir os parâmetros nesta etapa, por meio da qual, foi possível incorporar um mapa do Google ao PlaceProfile usando códigos JavaScript. Recursos de zoom e localização da cidade alvo, são funcionalidades aplicada ao PlaceProfile herdadas da API. A criação das células com as medidas precisas para largura e altura, levando em consideração a curvatura natural do planeta e a cor azul aplicada sobre o mapa, também são funcionalidades herdadas dessa API.

4.2 COLETA DOS DADOS

Nessa etapa os dados dados brutos sobre uma cidade ou região geográfica são coletados com o uso da biblioteca Google Place API. Na Tabela 3 é mostrado um exemplo de parte dos dados que podem ser coletados por meio de uma chamada a esta API. Os dados são extraídos em lotes, armazenados em um database e posteriormente processados. Diferentes tipos de processamento e análise podem ser realizados usando os mesmos dados. Na Figura 29 é apresentado o resultado da coleta dos POIs na área definida previamente na etapa de preparação, durante esse processo, os POIs coletados são plotados sobre a grid, trazendo a percepção ao usuário do volume de POIs vinculados a uma célula. Os pontos em vermelho são os POIs coletados em sua respectiva posição geográfica, obviamente que quanto mais intenso a quantidade de pontos vermelhos, maior a quantidade de POIs naquela célula. Na Figura 30 é aplicado o recurso de zoom sobre a área do mapa com 18 células para ilustrar os pontos coletados.

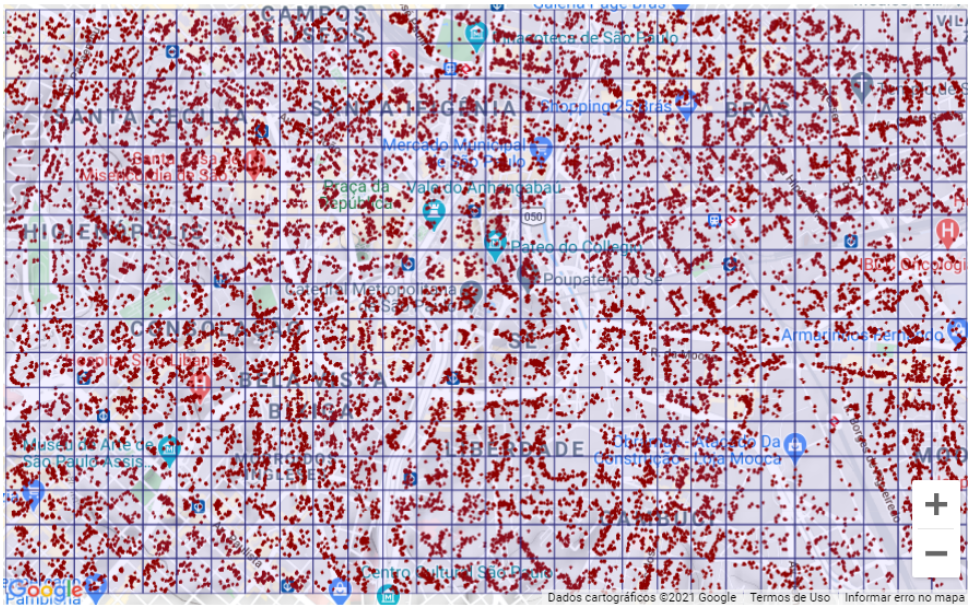

Figura 29 – PlaceProfile: Os dados coletados (pontos em vermelho) são plotados na grid em sua respectiva célula sobreposto ao mapa.

Tabela 3 – Dados brutos coletados do Google Place

Dados brutos sobre o POI	**Valores de exemplo**
ID	ChIJfUjHo85bzpQRjW2tr3wJs0k
Nome	Museu do Ipiranga
Coordenadas Geográficas	latitude : -23.5855993, longitude : -46.6097431
Tipo (130 possíveis valores)	museum, point_of_interest, establishment
Avaliaçào do usuário 1 (ruim) a 5 (bom)	4.6

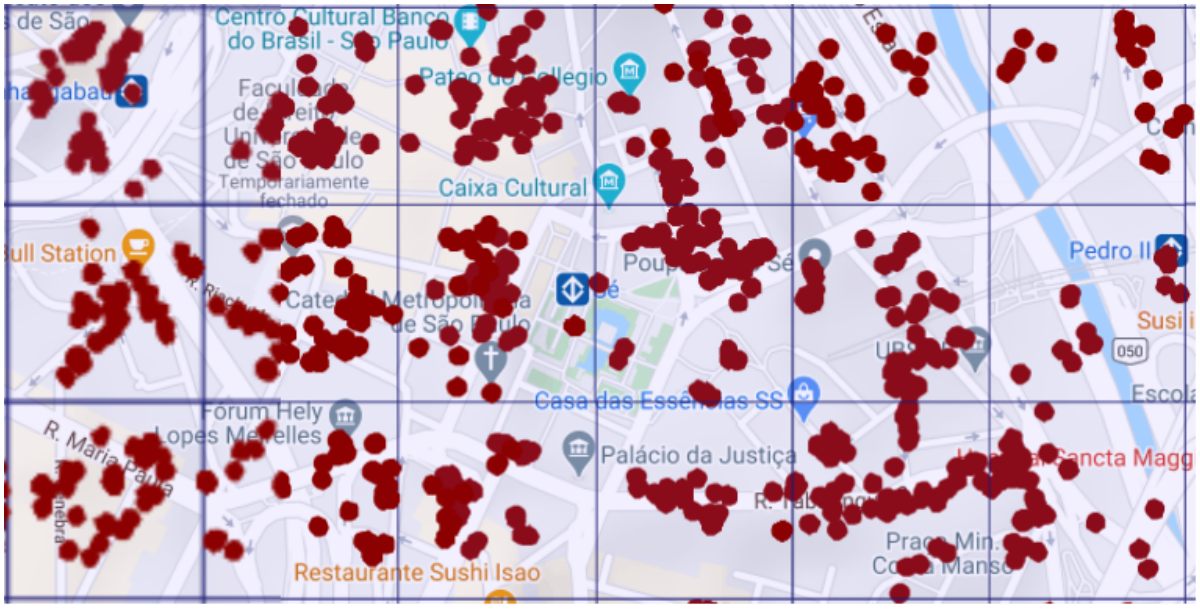

Figura 30 – PlaceProfile: Zoom aplicado sobre uma área de coleta com 18 células selecionadas.

4.3 MINERAÇÃO DOS DADOS

A etapa de mineração é a etapa em que todo o processo ocorre em backend antes de retornar a informação para o módulo frontend do PlaceProfile para a visualização das dados produzidos. Primeiramente todos os 130 tipos de categorias identificados do Google Places são agrupados em 11 macrocategorias, conforme mostrado na Tabela 4, possibilitando o usuário selecionar ou descartar macrocategorias para análise. Essa mesma técnica de agrupamento das categorias em macrocategorias foi realizado anteriormente no trabalho de (D'Andrea et al., 2018). Em seguida, é realizada a contagem de macrocategorias por célula criando novos atributos, como mostrado na Tabela 5. Esses novos atributos serão passados como parâmetros para o algoritmos de agrupamento e para a abordagem de visualização iconográfica. A atual implementação do PlaceProfile inclui somente os algoritmos de agrupamento como o K-means, C-means e Agglomerative para definir o perfil das áreas da cidade, podendo posteriormente ser aplicado diferentes algoritmos de machine learning para extrair informações úteis dos dados disponíveis. O usuário pode ainda incluir outros atributos para análise como as coordenadas geográficas, identificador da célula ou número da linha e da coluna da célula na grid. Células que possuem poucos POIs, também podem ser descartadas da análise, isso a critério do usuário. Por fim, como parâmetro obrigatório, a quantidade de grupos que se deseja dividir a base de dados, ou seja, quantas k áreas distintas sobre os dados serão rotuladas para criação do perfil dessa região.

Os dados resultantes do processo de mineração são novamente armazenados em um database para serem usados na etapa visualização ou exportados para uso em outras

Tabela 4 – Tabela relacionando as diversas categorias de atividades similares capturadas do Google Places API, com suas respectivas macrocategorias, essa técnica foi usado no trabalho de D'Andrea et al. (2018).

Macrocategoria	**Lista de categorias originadas do Google Maps**
Food	bakery, bar, cafe, food, liquor_store, meal_delivery, meal_takeaway, restaurant
Finance	accounting, atm, bank, finance
Administrative	city_hall, courthouse, embassy, fire_station, local_government_office, police
Transportation	airport, bus_station, subway_station, taxi_stand, train_station, transit_station, light_rail_station
Cultural	art_gallery, library, school, university, movie_theater, museum
Entertainment	night_club, amusement_park, casino, bowling _lley, campground, zoo, aquarium, stadium
Health	pharmacy, physiotherapist, beauty_salon, dentist, doctor, gym, hair_care, hospital, veterinary_care, health, spa
Services	travel_agency, funeral_home, painter, park, post_office, parking, roofing_contractor, locksmith, general_contractor, lodging, moving_company, car_repair, car_wash, electrician, car_rental, laundry, gas_station, plumber, real_estate_agency, recreational_vehicle_park, insurance_agency, lawyer
Religious	mosque, cemetery, church, hindu_temple, synagogue, place_of_worship
Stores	shoe_store, shopping_mall, pet_store, bicycle_store, book_store, car_dealer, clothing_store, jewelry_store, florist, store, furniture_store, convenience_store, department_store, electronics_store, hardware_store, home_goods_store, storage, grocery_or_supermarket, movie_rental
Miscellaneous	point_of_interest, establishment, country, floor, intersection, locality, natural_feature, geocode, colloquial, area, room, post_box, neighborhood, postal_code, postal_town, political, postal_code_prefix, postal_code_suffix, premise, route, street_address, subpremise, street_number, sublocality_(SL)

aplicações que desejam usar esses dados em suas pesquisas.

Tabela 5 – Uma amostra da sumarização das macrocategorias por grupos(clusters): a contagem do total das 11 macrocategorias em cada célula cria os atributos que serão passados como parâmetro para o algoritmo de agrupamento e para a análise iconográfica .

		Contagem de Macrocategorias				
Id	**Coordenadas(lat, lng)**	**Food**	**Finance**	**Admin**	**Transport**	**Cultural**
1	-23.532288, -46.671019	5	10	3	3	5
2	-23.532288, -46.668570	14	1	0	2	2
3	-23.532288, -46.666120	16	3	1	1	6
4	-23.532288, -46.663671	6	9	4	2	0

4.4 VISUALIZAÇÃO

Na etapa de visualização, três diferentes estratégias são usadas para representar os resultados das etapas anteriores: (i) Análise Iconográfica, (ii) Análise por agrupamento e (iii) Análise usando o gráfico Radar Chart.

A biblioteca Maps Javascript API do Google foi utilizada para a renderização do gráfico sobre o mapa e para a coloração de células baseados nos resultados das análise. A seguir é detalhado a construção das visualizações usando recursos gráficos dessa API.

Para o desenvolvimento da análise iconográfica cada célula foi dividida em 100 partes iguais (veja na Figura 31), criando uma nova grid interna de tamanho de 10x10 (a). Em cada célula são destacadas as 4 principais macrocategorias naquela célula, ou seja, as macrocategorias que mais se repetem proporcionalmente dentro da célula. Para isso, a macrocategorias de POIs de maior destaque aparece no topo da célula, em seguida a segunda macrocategoria em destaque e a terceira macrocategoria em destaque vem logo a seguir. Por fim, no restante das células da grid interna, aparece a soma de todas as outras macrocategorias identificadas e presentes nessa célula (b).

A Figura 32 ilustra um exemplo da visualização dos perfis por meio da abordagem iconográfica empregada no mesmo resultado da divisão das células e suas principais atividades apresentados na Figura 31. Observe que cada cor representa uma macrocategoria e a ocupação da cor em cada célula representa a proporção de atividades relacionadas a essa macrocategoria. Neste exemplo, a cor roxa representa a macrocategoria de serviços (*services*), o vermelho está relacionado às lojas (*stores*), o amarelo representa alimentação (*food*), o verde representa a saúde (*health*), o azul representa cultural (*cultural*) e o marrom representa a soma de todas as outras macrocategorias presentes na célula. O usuário pode notar as categorias que não são predominantes em cada célula observando a legenda que descreve cada macrocategoria empregada na visualização.

Na visualização de análise de cluster cada célula é rotulada com uma cor diferente para cada grupo de dados particionado pelo algoritmo. A Figura 33 mostra o resultado da

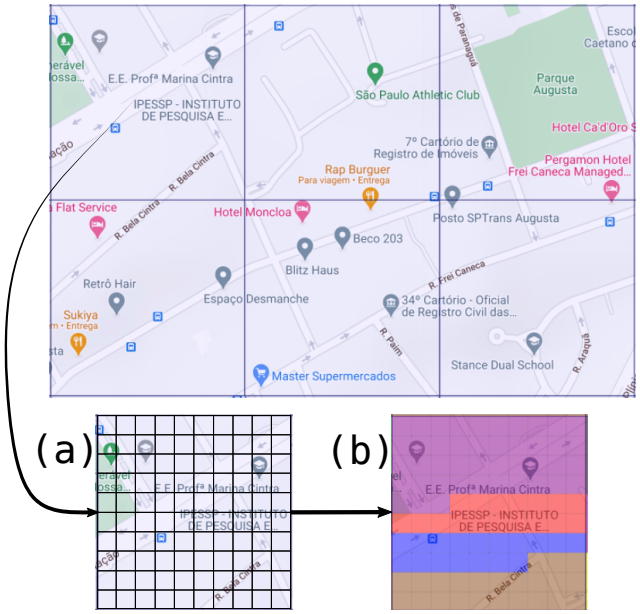

Figura 31 – PlaceProfile: Ilustração da técnica de visualização aplicada para destacar as mais comuns macrocategorias por célula. Uma grid é sobreposta a uma região e cada célula da grid é dividida novamente em uma outra subgrid 10x10 (a) e as macrocategorias mais presentes nessa célula são ordenadas de cima para baixo de acordo com a quantidade de atividades relacionadas a macrocategoria (b). A última cor(marrom), está relacionada a soma de todas as outras macrocategorias identificadas na célula.

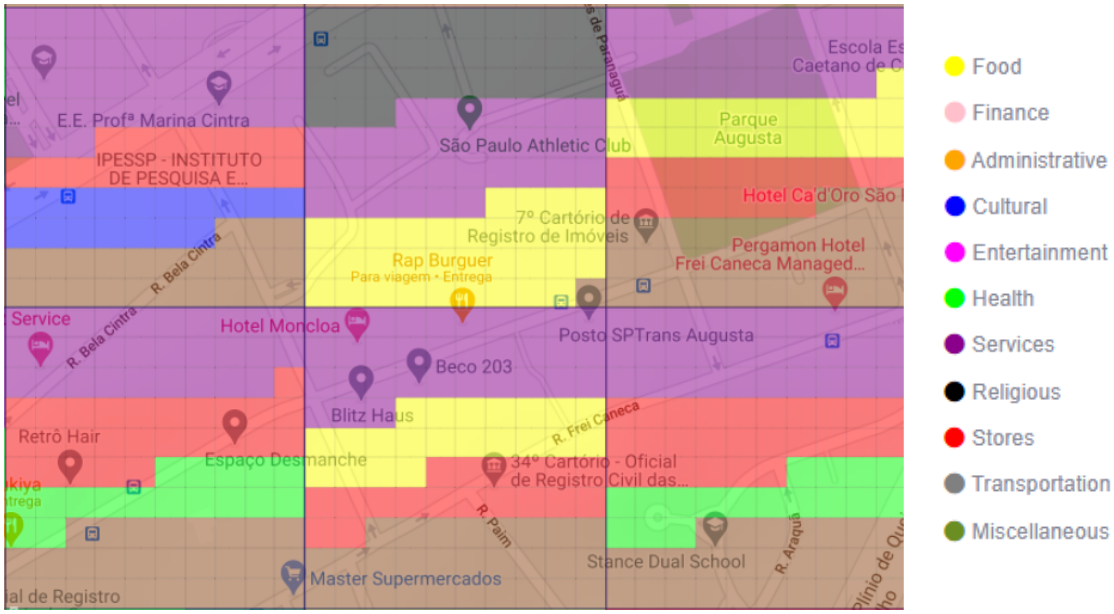

Figura 32 – PlaceProfile: Usando cores para codificar a macrocategoria mais comum em uma célula. Esta região é particularmente representada por macrocategorias de serviços (*services*) e lojas (*stores*).

análise de cluster nas mesmas seis células destacadas nas Figuras 31 e 32. Observe que cinco das seis células pertencem ao mesmo cluster (células em vermelho) e uma célula tem características diferentes (célula em verde). Devemos destacar que as cores na análise de

cluster não se relacionam com as cores utilizadas para codificar as macrocategorias.

Para complementar a análise das macrocategorias e auxiliar na interpretação dos resultados dos clusters gerados, o PlaceProfile usa o gráfico Radar Chart em coordenação com a visualização de análise dos clusters e a visualização. Isso permite que os usuários a percebam as macrocategorias predominantes para cada célula. Por meio do Radar Chart o usuário pode ter uma melhor percepção de como as informações na célula levaram a formação do cluster, mostrando a proporção de POIs para cada cluster. A proporção de cada macrocategoria no cluster é calculada somando a quantidade de uma respectiva macrocategoria dividido pela soma total de macrocategorias no cluster.

A interatividade está presente nesta visualização, permitindo ao usuário selecionar células de interesse para analisar e fazer comparações. A Figura 34 mostra as diferenças entre os dois clusters, as células vermelhas se destacam nas macrocategorias de serviços (*services*) e atividades de lojas (*stores*), enquanto na célula verde o destaque é para lojas (*stores*), serviços (*services*) e saúde (*health*).

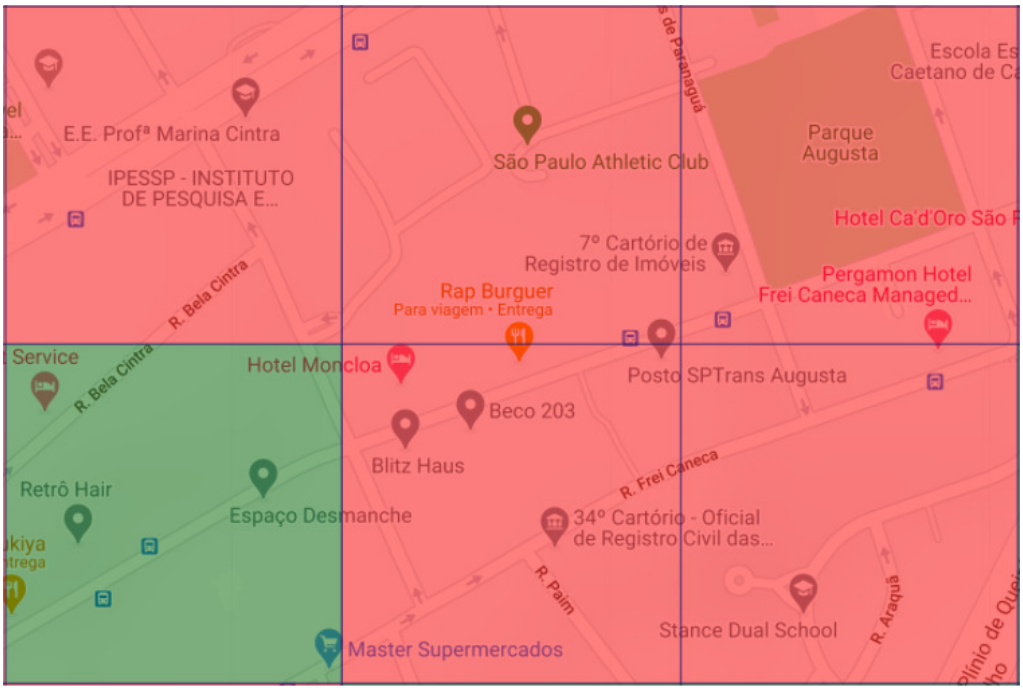

Figura 33 – PlaceProfile: Análise de cluster. Com base nos recursos recuperados durante a etapa de coleta de dados, as células são agrupadas para ajudar na análise com base na similaridade.

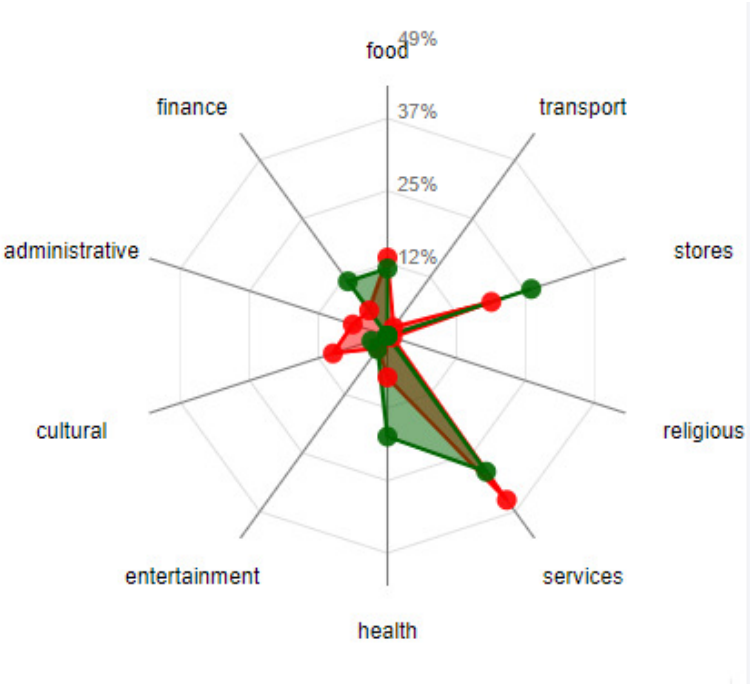

Figura 34 – PlaceProfile: O gráfico Radar Chart mostra a proporção de macrocategorias para os dois clusters apresentados na Figura 33. Embora muito semelhantes em serviços (*services*), lojas (*stores*) e alimentação (*food*), esses dois grupos diferem nas macrocategorias cultural (*cultural*), de saúde *health* e financeira (*finance*).

4.5 IMPLEMENTAÇÃO

Para o desenvolvimento do frontend do PlaceProfile foi usando a linguagem de marcação de hipertextos HTML5, para estilização do site foi usado a linguagem CSS3 e para a execução das ações de interatividade do web site foi usado a linguagem Javascript. Além do mais, foi usado as bibliotecas Bootstrap[1] para otimizar o processo de estilização e a biblioteca JQuery[2] para manipulação de eventos e interatividade com APIs externas. No desenvolvimento do backend do PlaceProfile, a linguagem usada foi Python[3], junto com as bibliotecas pandas[4] para o processo de limpeza e mineração dos dados brutos e scikit-learn[5] para a aplicação dos algoritmos de agrupamento.

[1] https://getbootstrap.com/
[2] https://jquery.com/
[3] https://www.python.org/
[4] https://pandas.pydata.org/
[5] https://scikit-learn.org/

5 RESULTADOS

Para demonstrar o uso do PlaceProfile, este capítulo apresentada os resultados da análise sobre os POIs coletados sobre região central da cidade de São Paulo/Brasil. Nessa amostra, foi definido que cada célula terá o tamanho de 250x250 metros, gerando assim uma grid com 17 linhas e 28 colunas, totalizando 476 células sobreposta à cidade de São Paulo. Como resultado, foram coletados e armazenados 25302 POIs. Na Figura 35 é apresentada a região antes da coleta dos dados com a marcação da grid sobrepostas ao mapa, e na Figura 36 é apresentada a mesma região já com a coleta dos dados realizado.

Figura 35 – PlaceProfile: A grid plotada sobre o mapa de uma área da cidade de São Paulo, cada célula da grid tem um tamanho de 250x250 metros, totalizando 476 células.

5.1 ANÁLISE VISUAL DOS PERFIS

Para a análise visual dos perfis de cada região, foram selecionados 10 de 11 macrocategorias para análise, desconsiderado apenas a categoria *miscellaneous* devido a não relação dessa macrocategoria a nenhum outro tipo de macrocategoria. Na Figura 37, por meio da análise visual dos perfis, é possível observar as macrocategorias de maior destaque em cada célula. Por exemplo, na região mais ao oeste da área de São Paulo visível no mapa, as macrocategorias em destaque são relacionados a área da saúde (*health*) em verde; já mais ao nordeste do mapa, as macrocategorias em destaque são relacionadas a lojas (*stores*) em vermelho, outras células também possuem macrocategorias diferentes em destaque como alimentação (*food*) em amarelo e serviços (*service*)em roxo. As células na cor preta

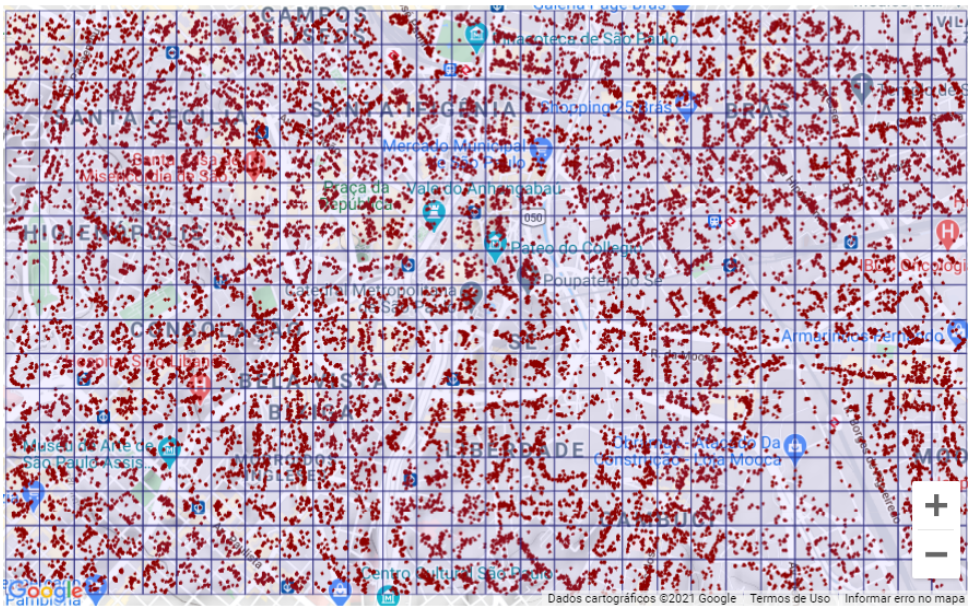

Figura 36 – PlaceProfile: Os dados coletados (pontos em vermelho) são plotados na grid em sua respectiva célula e posição geográfica sobreposto ao mapa.

são células que foram descartadas por não terem a quantidade mínima de POIs, neste caso 1.

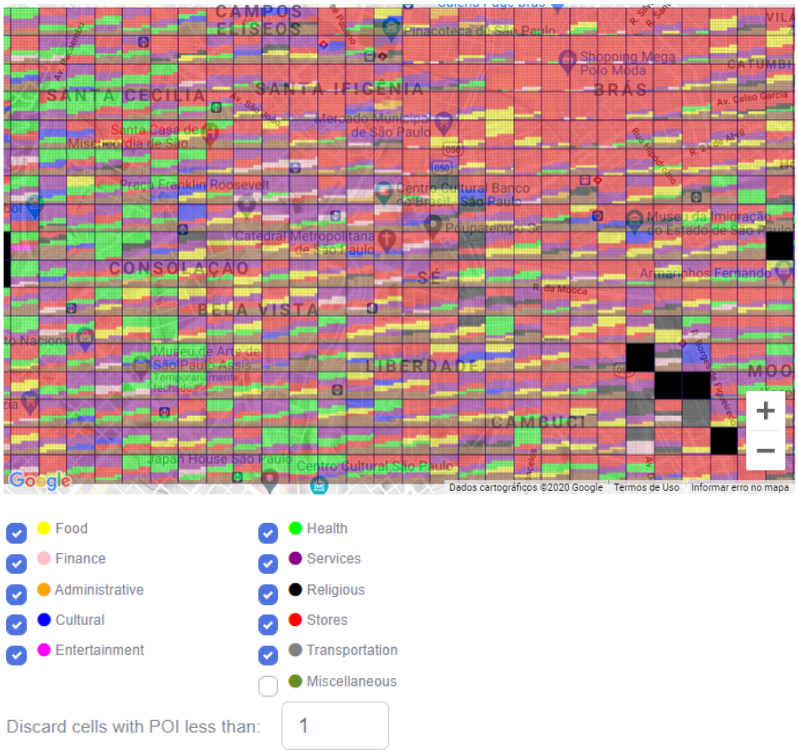

Figura 37 – PlaceProfile: Uma visão geral sobre a região central da cidade de São Paulo. Podemos ver principalmente uma divisão entre POIs relacionados a lojas (*stores*) e POIs relacionados a saúde (*health*). As células pretas correspondem a regiões com número de POIs abaixo do limite mínimo definido pelo usuário.

5.2 ANÁLISE VISUAL DOS ALGORITMOS DE AGRUPAMENTO

A seguir, é realizada a análise da similaridade entre os POIs coletados para esta região aplicando análise de agrupamento. Usando os mesmos dados e os mesmos filtros aplicados nas etapas de pré-processamento (preparação e coleta dos dados), discutidas na seção anterior, foram selecionados 10 de 11 macrocategorias para análise, desconsiderado apenas a categoria *miscellaneous* devido a não relação dessa macrocategoria a nenhum outro tipo de macrocategoria.

Atualmente, o PlaceProfile oferece como opção ao usuário 3 tipos de algoritmos de agrupamento, o K-means (KRISHNA; MURTY, 1999), o Agglomerative (KURITA, 1991) e o Fuzzy C-means (CANNON; DAVE; BEZDEK, 1986). Para cada um dos experimentos, foi definido k=4, mais uma classe de células que serão descartadas da análise por não ter a quantidade mínima de POIs, neste caso definido como 1, totalizando 5 classes. Assim, todas as células que não tiverem o mínimo de 1 ponto de interesse, serão desconsiderados da análise. Esse descarte é opcional ao usuário e caso não opte pelo descarte o número de classes será a mesma que foi definido para k.

Nesse experimento, os algoritmos foram aplicados no espaço de alta dimensionalidade composto pelos atributos (características), sem utilizar as coordenadas de latitude e longitude, conforme indicado na Tabela 4 e sumarizados como na Tabela 5.

Na Figura 38 permite fazer um comparativo visual entre a análise iconográfica e os algoritmos de agrupamento. É possível observar que os 3 algoritmos apresentaram resultados similares, representando em seus clusters áreas muito similares a aquelas também identificadas na análise iconográfica, principalmente em áreas onde uma ou duas macrocategorias se destacam mais em relação a quantidade de atividades na célula, a exceção se aplica ao algoritmo Agglomerative que destoou dos demais generalizando em um mesmo cluster (azul) grupos que os outros algorítimos classificaram em grupos separados.

Um dos problemas de analisar o resultado de um algoritmo de agrupamento é que as informações sobre quais atributos (características) levaram aos padrões de agrupamento geralmente são perdidas. Para tentar explicar como as informações nas regiões levaram à formação de clusters, a análise visual para cada algoritmo de agrupamento é apoiada ao gráfico Radar Chart. O objetivo é mostrar a proporção de POIs em cada região por meio de um mecanismo de coordenação. As Figuras 39, 40 e 41 apresentam; (a) os resultados visuais aplicados aos algoritmos de agrupamento; (b) apoiado pelo gráfico Radar Chart passado a ideia da exata da proporção de macrocategorias por cluster.

É possível perceber que o Radar Chart codifica a proporção de POIs para as células. Na análise de similaridade baseada em resultados de agrupamento usando o algoritmo K-means (Figura 39), é possível saber que o cluster azul representa os POIs relacionados à macrocategoria loja (*store*), enquanto o cluster em laranja representa os POIs relacionados

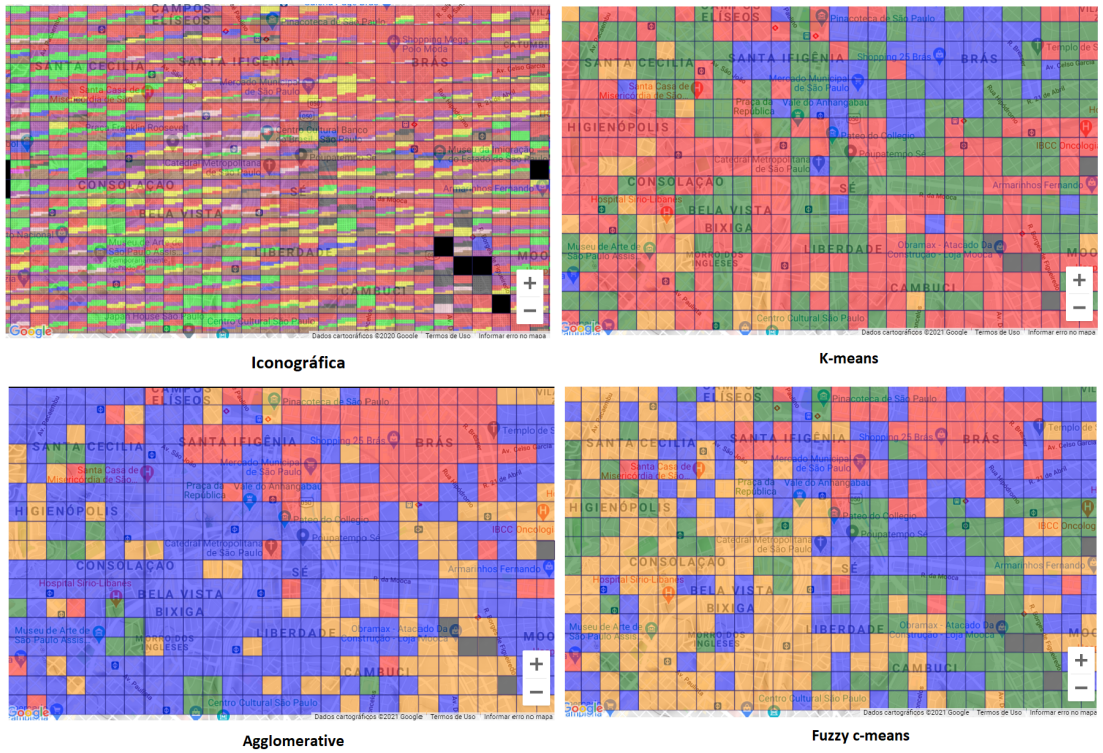

Figura 38 – PlaceProfile: Um comparativo do resultado visual demonstrando os resultados da análise iconográfica e para cada um dos algoritmos de agrupamento. Nesse cenário, é possível perceber que o Agglomerative destoou dos demais generalizando em um mesmo cluster grupos que os outros algoritmos classificaram em grupos separados.

às macrocategorias de saúde (*health*), já as células em vermelho não há uma macrocategoria com grande destaque sobre as outras. Na análise de similaridade baseada em resultados de agrupamento usando o algoritmo Agglomerative (Figura 40), o cluster vermelho representa os POIs relacionados à macrocategoria loja (*store*), enquanto o cluster verde representa os POIs relacionados às macrocategorias de saúde (*health*), já as células em azul e laranja não há uma macrocategoria com grande destaque sobre as outras. E, por fim, na análise de similaridade baseada em resultados de agrupamento usando o algoritmo Fuzzy C-means (Figura 41), os clusters nas cores vermelho e azul representam os POIs relacionados à macrocategoria loja (*store*), porém o cluster azul também há uma incidência de atividades relacionados à macrocategoria serviços (*services*) e saúde (*health*), enquanto o cluster amarelo representa os POIs relacionados às macrocategorias de saúde (*health*), serviços (*services*) e lojas (*stores*), já as células em verde não há uma macrocategoria com grande destaque sobre as outras.

Observe que esta análise é consistente com a representação iconográfica vista na Figura 37. Ou seja, enquanto as células que estão relacionadas à região nordeste da Figura 37, e possuem em destaque à macrocategoria loja(*stores*), são as mesmas células que também são destacadas nas Figuras 39, 40 e 41 como lojas (*stores*), e o mesmo é

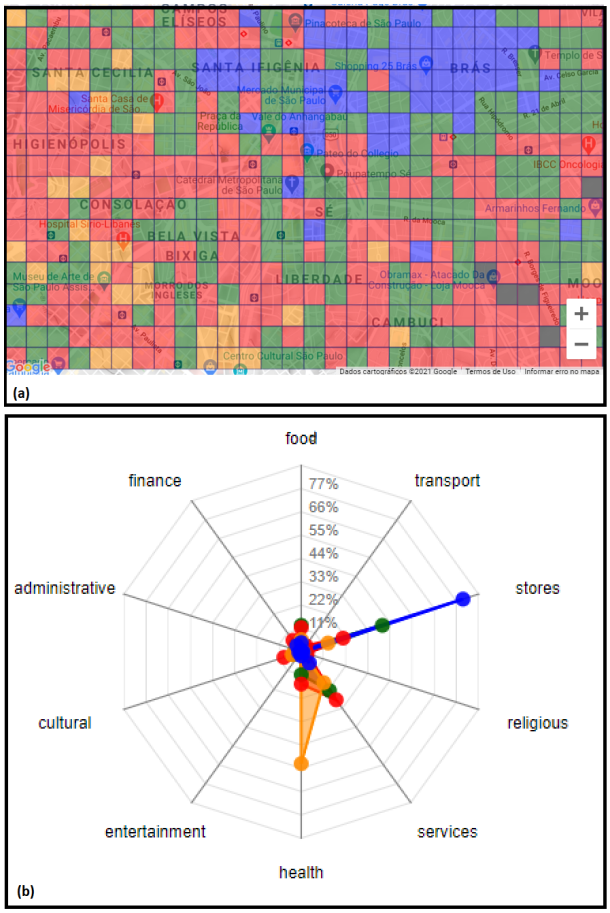

Figura 39 – PlaceProfile: Análise de similaridade baseada em resultados de agrupamento usando o algoritmo K-means. O Radar Chart apoia na visualização mostrando que o cluster azul representa os POIs relacionados à macrocategoria loja (*store*), enquanto o cluster laranjas representa os POIs relacionados às macrocategorias de saúde (*health*), já as células em vermelho não há uma macrocategoria com grande destaque sobre as outras.

confirmado na região mais ao oeste que é consistente com as células mostrando POIs relacionados à saúde (*health*) e serviço (*service*).

5.2.1 ANÁLISE A PARTIR DE CÉLULAS SELECIONADAS

O PlaceProfile permite ao usuário selecionar quais células ele deseja analisar e fazer comparações no Radar Chart. Na Figura 42, foi aplicado a análise de agrupamento usando o algoritmo K-means. Foi possível filtrar as células para análise identificando a proporção de POIs para as células vermelhas e azuis, selecionadas e destacadas na visualização da grid com bordas mais grossas. Assim, é possível saber como aquelas células selecionadas pertencentes ao cluster azul apresentam uma proporção muito maior de POIs relacionados à macrocategoria lojas (*stores*) do que qualquer outra, enquanto que as células selecionadas do cluster vermelho apresentam POIs relacionados às macrocategorias de serviços (*services*) e saúde (*health*). Observe que esta análise é consistente com a representação iconográfica

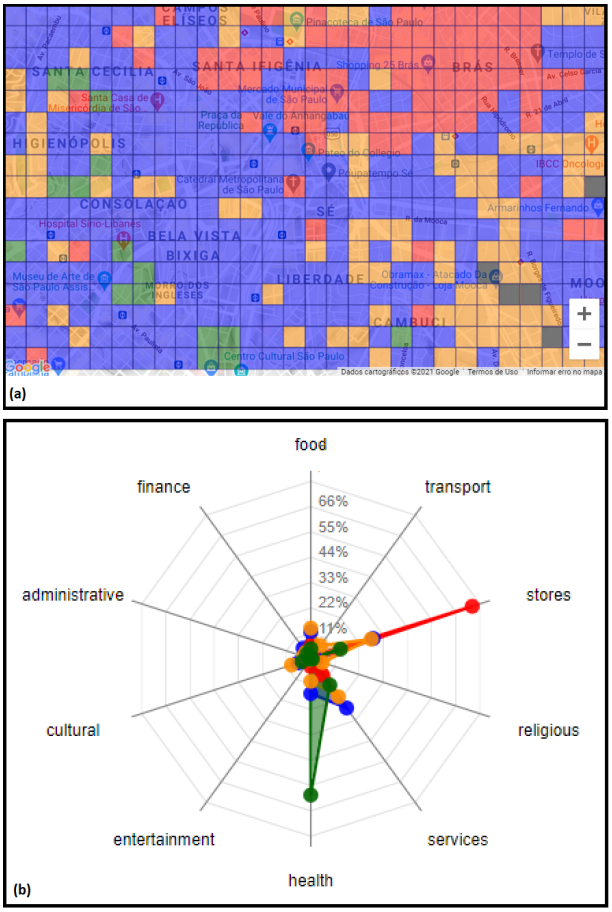

Figura 40 – PlaceProfile: Análise de similaridade baseada em resultados de agrupamento usando o algoritmo Agglomerative. O Radar Chart apoia na visualização mostrando que o cluster vermelho representa os POIs relacionados à macrocategoria loja (*store*), enquanto o cluster verde representa os POIs relacionados às macrocategorias de saúde (*health*), já as células em azul e laranja não há uma macrocategoria com grande destaque sobre as outras.

vista na Figura 37. Ou seja, enquanto o cluster azul está relacionado à região nordeste da Figura 37 – células com POIs altamente relacionados macrocategoria lojas, o cluster vermelho é consistente com as células mostrando POIs relacionados à saúde e serviços na mesma figura.

A Figura 43 mostra o mesmo resultado de agrupamento, mas com diferentes células selecionadas. Usando a coordenação entre o mapa e o gráfico de radar para ajudar na explicabilidade do cluster. Usando o Radar Chart, as células do cluster verde representam POIs relacionados à macrocategorias lojas (*stores*) e serviços (*services*). É possível observar que, essas células são aquelas localizadas nos limites do cluster azul e na vizinhança da área concentrada com POIs da macrocategoria loja (*stores*) na Figura 37. Por último, vemos no Radar Chart que o cluster amarelo corresponde aos POIs relacionados à macrocategoria saúde (*health*). Esta informação também é percebida usando a representação iconográfica da Figura 37.

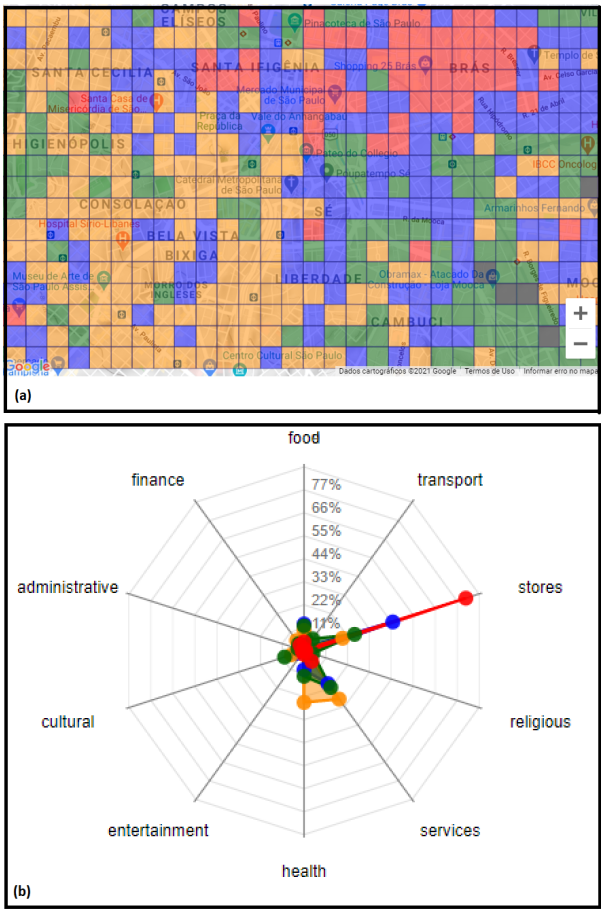

Figura 41 – PlaceProfile: Análise de similaridade baseada em resultados de agrupamento usando o algoritmo Fuzzy C-means. O Radar Chart apoia na visualização mostrando que os clusters na cor vermelho e azul representam os POIs relacionados à macrocategoria loja (*store*), porém o cluster azul também há uma incidência de atividades relacionados à macrocategoria serviços (*services*) e saúde (*health*), enquanto o cluster amarelo representa os POIs relacionados às macrocategorias de saúde (*health*), serviços (*services*) e lojas (*stores*), já as células em verde não há uma macrocategoria com grande destaque sobre as outras.

Neste estudo de caso, foi demonstrado como o mecanismo de explicabilidade do Radar Chart pode ajudar os usuários a entender a formação do cluster a partir da análise da proporção de POIs nas células dos clusters de interesse. No entanto, a abordagem iconográfica oferece o mesmo poder de análise, pois codifica padrões de células que apresentam proporção semelhante de POIs com a mesma categoria. A abordagem iconográfica pode fornecer uma visão geral da organização dos dados e detalhes sobre os POIs ao mesmo tempo.

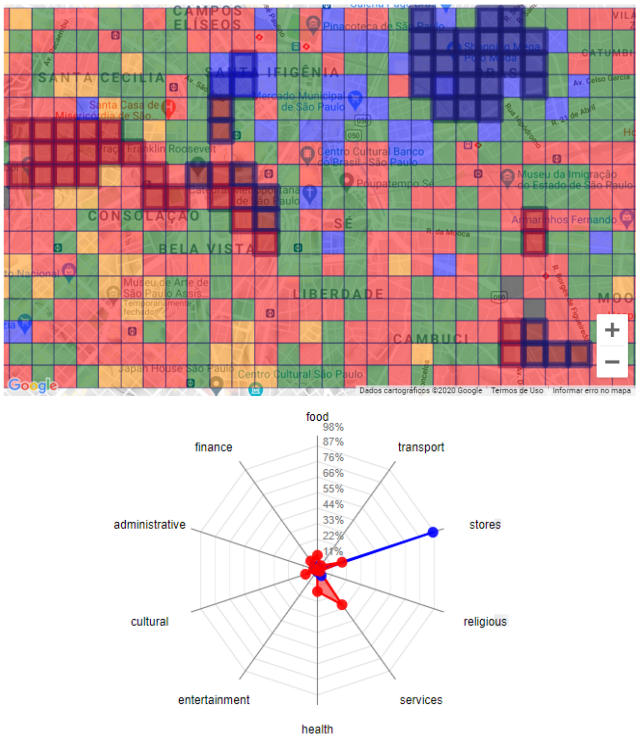

Figura 42 – PlaceProfile: Análise de similaridade baseada em resultados de agrupamento. As células selecionadas mostram que o cluster azul representa os POIs relacionados à macrocategoria loja (*store*), enquanto o cluster vermelho há um pequeno destaque para serviços (*service*), mas também há incidências de outras macrocategorias como de saúde (*health*), lojas (*stores*).

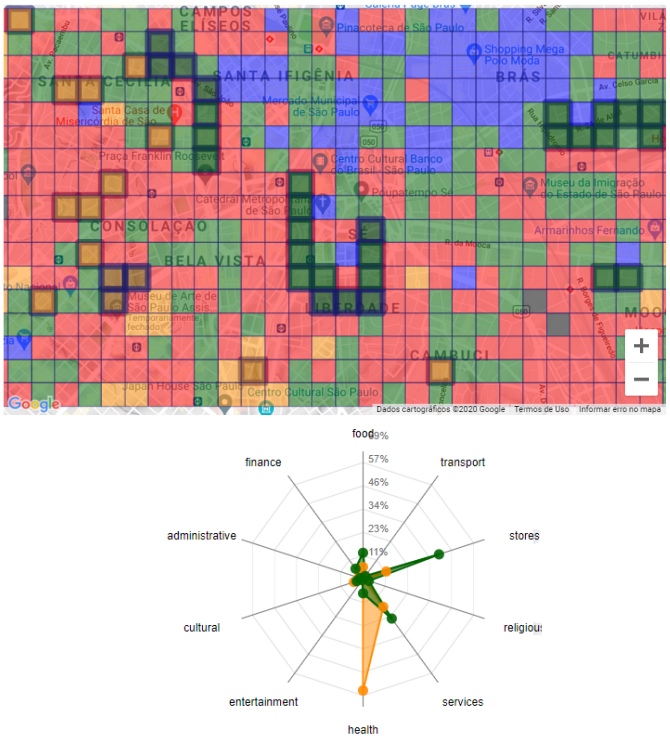

Figura 43 – PlaceProfile: Seleção de diferentes clusters para entender os padrões de POIs. O cluster verde parece ter POIs relacionados às macrocategorias de lojas (*stores*) e serviços (*services*), enquanto o cluster amarelo corresponde aos POIs relacionados à macrocategoria de saúde (*health*).

6 CONCLUSÃO E TRABALHOS FUTUROS

Este trabalho apresentou o PlaceProfile, uma ferramenta de visualização baseada na web para identificar o perfil de áreas em cidades ou regiões metropolitanas. O PlaceProfile permite a definição de regiões, granularidade de análise e outros parâmetros para auxiliar na análise de padrões com base em pontos de interesse. Outras pesquisas já realizaram trabalhos similares com objetivo de rotular áreas em uma região (D'Andrea et al., 2018), (Noulas; Mascolo; Frias-Martinez, 2013), mas, de acordo com a revisão bibliográfica realizada, a análise realizada nesses trabalhos ficaram limitadas a uma única cidade.

A principal vantagem do PlaceProfile consiste na capacidade de fornecer o conhecimento gerado pelo usuário. Ele fornece informações e mecanismos de exploração para auxiliar analistas a gerar conhecimento sobre as áreas em análise. Primeiramente, os resultados do agrupamento foi aumentado coordenando um mapa com um Radar Chart que mostra a proporção de pontos de interesse nas células selecionadas. Assim, os usuários podem conhecer como essas células se diferenciam ou se relacionam para contribuir com a formação dos agrupamentos. Em segundo lugar, nossa abordagem iconográfica estende a análise de cluster, mostrando uma visão geral e informações detalhadas ao mesmo tempo. Em um nível superior, os usuários entendem o resultado e a possível formação de agrupamentos inspecionando os padrões de cores usados para o design iconográfico. Em uma análise detalhada, os usuários inspecionam a proporção de pontos de interesse dentro das células do cluster. Além disso, o PlaceProfile permite ao usuário maior interatividade, possibilitando a filtragem de macrocategorias e análise de células selecionadas em uma região.

Como estudo de caso, foi aplicado a análise dessas duas abordagens sobre dados amostrais coletados de POIs na cidade de São Paulo/Brasil, mas em experimentos realizados, o PlaceProfile se mostrou funcional também para outras cidades. A abordagem ajudou a entender como os pontos de interesse estão organizados de forma que áreas próximas apresentem categorias semelhantes. Além disso, o estudo de caso também demonstrou como as duas estratégias de análise (clustering e iconográfica) são consistentes uma com a outra.

Durante as etapas de preparação e coleta dos dados, é necessário que o usuário faça experimentos de coleta e análise algumas vezes, alterando o tamanho das células e aplicando diferente perspectivas de zoom sobre o mapa da região que se deseja analisar até ele concluir que o resultado está satisfatório. Foi usado algoritmos de agrupamento para separar o conjunto de dados, mas é possível ser aplicado diferentes algoritmos de machine learning para extrair informações úteis dos dados disponíveis. Outra fontes de dados

também podem ser incluídas nessas análises como quantidade de check-ins que usuários de aplicativos de redes sociais fazem em um local especifico e avaliações de usuários com intuito de analisar a satisfação do local. Mais uma etapa de interação com o usuário pode ser adicionado com o objetivo de permitir ao usuário que adicione a determinadas atividade em uma região pesos diferentes, influenciado a análise do algoritmo, por exemplo, no atual cenário do PlaceProfile, uma universidade que tem muitos alunos e funcionários tem o mesmo peso de uma pequena papelaria situada na mesma célula analisada.

Em trabalhos futuros e com base nos dados produzidos neste trabalho é interessante combinar dados sobre deslocamento humano, identificando o perfil de áreas em que grupos de pessoas se deslocam com intuito de identificar o que leva essas pessoas a se deslocarem e recomendar opções de atividades em que o usuário habitualmente realiza, de acordo com o cluster e o perfil iconográfico, em locais mais próximos ao seu ponto de origem.

REFERÊNCIAS

ALBINO, V.; BERARDI, U.; DANGELICO, R. M. Smart cities: Definitions, dimensions, performance, and initiatives. *Journal of Urban Technology*, Routledge, v. 22, n. 1, p. 3–21, 2015. Citado na página 37.

ANDRIENKO, G. et al. Visual analytics focusing on spatial events. In: *Visual analytics of movement*. [S.l.]: Springer, 2013. p. 209–251. Citado 3 vezes nas páginas x, 3 e 34.

ASTEL, A. et al. Clasification of drinking water samples using the chernoff's faces visualization approach. *Polish Journal of Environmental Studies*, v. 15, n. 5, 2006. Citado 2 vezes nas páginas ix e 19.

BACKSTROM, L.; SUN, E.; MARLOW, C. Find me if you can: Improving geographical prediction with social and spatial proximity. In: *Proceedings of the 19th International Conference on World Wide Web*. New York, NY, USA: ACM, 2010. (WWW '10), p. 61–70. ISBN 978-1-60558-799-8. Disponível em: <http://doi.acm.org/10.1145/1772690.1772698>. Citado na página 26.

BAO, F.; CHEN, J. Visual framework for big data in d3.js. In: IEEE. *2014 Ieee Workshop on Electronics, Computer and Applications*. [S.l.], 2014. p. 47–50. Citado na página 16.

BATTY, M. Cities as complex systems: Scaling, interaction, networks, dynamics and urban morphologies. Springer, 2009. Citado na página 3.

BIELZA, C.; LARRAÑAGA, P. Discrete bayesian network classifiers: a survey. *ACM Computing Surveys (CSUR)*, ACM, v. 47, n. 1, p. 5, 2014. Citado na página 8.

BOWMAN, J. L.; BEN-AKIVA, M. E. Activity-based disaggregate travel demand model system with activity schedules. *Transportation research part a: policy and practice*, Elsevier, v. 35, n. 1, p. 1–28, 2001. Citado 3 vezes nas páginas x, 30 e 31.

CANNON, R. L.; DAVE, J. V.; BEZDEK, J. C. Efficient implementation of the fuzzy c-means clustering algorithms. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, n. 2, p. 248–255, 1986. Citado 2 vezes nas páginas 9 e 52.

CHEN, M. et al. Big data: related technologies, challenges and future prospects. Springer, 2014. Citado na página 7.

CHEN, S. et al. Interactive visual discovering of movement patterns from sparsely sampled geo-tagged social media data. *IEEE transactions on visualization and computer graphics*, IEEE, v. 22, n. 1, p. 270–279, 2015. Citado 3 vezes nas páginas xi, 37 e 39.

CHEN, Y.-C. et al. Interactive visual analysis for vehicle detector data. In: WILEY ONLINE LIBRARY. *Computer Graphics Forum*. [S.l.], 2015. v. 34, n. 3, p. 171–180. Citado na página 35.

CHO, E.; MYERS, S. A.; LESKOVEC, J. Friendship and mobility: user movement in location-based social networks. In: ACM. *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. [S.l.], 2011. p. 1082–1090. Citado na página 26.

CLARAMUNT, C.; JIANG, B.; BARGIELA, A. A new framework for the integration, analysis and visualisation of urban traffic data within geographic information systems. *Transportation Research Part C: Emerging Technologies*, Elsevier, v. 8, n. 1-6, p. 167–184, 2000. Citado 3 vezes nas páginas ix, 3 e 22.

D'Andrea, E. et al. Smart profiling of city areas based on web data. In: *2018 IEEE International Conference on Smart Computing (SMARTCOMP)*. [S.l.: s.n.], 2018. p. 226–233. Citado 14 vezes nas páginas ix, x, xiv, 1, 2, 3, 7, 23, 24, 26, 27, 44, 45 e 59.

DATAVIZCATALOGUE. *The Data Visualisation Catalogue Blog*. 2021. Disponível em: <http://datavizcatalogue.com/blog/chart-combinations-tile-grid-maps/>. Citado 2 vezes nas páginas ix e 19.

DEFAYS, D. An efficient algorithm for a complete link method. *The Computer Journal*, Oxford University Press, v. 20, n. 4, p. 364–366, 1977. Citado na página 15.

DEMISSIE, M. G.; CORREIA, G. H. de A.; BENTO, C. Exploring cellular network handover information for urban mobility analysis. *Journal of Transport Geography*, Elsevier, v. 31, p. 164–170, 2013. Citado 4 vezes nas páginas x, 3, 36 e 38.

DUNN, J. C. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. Taylor & Francis, 1973. Citado na página 13.

ESTER, M. et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In: *Kdd*. [S.l.: s.n.], 1996. v. 96, n. 34, p. 226–231. Citado na página 29.

ESTIVILL-CASTRO, V. Why so many clustering algorithms: A position paper. *SIGKDD Explor. Newsl.*, ACM, New York, NY, USA, v. 4, n. 1, p. 65–75, jun. 2002. ISSN 1931-0145. Disponível em: <http://doi.acm.org/10.1145/568574.568575>. Citado na página 8.

GANDOMI, A.; HAIDER, M. Beyond the hype: Big data concepts, methods, and analytics. *International journal of information management*, Elsevier, v. 35, n. 2, p. 137–144, 2015. Citado na página 10.

GOSAIN, A.; BHUGRA, M. A comprehensive survey of association rules on quantitative data in data mining. In: IEEE. *2013 IEEE Conference on Information & Communication Technologies*. [S.l.], 2013. p. 1003–1008. Citado na página 9.

HASTIE, T. et al. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, Springer, v. 27, n. 2, p. 83–85, 2005. Citado na página 31.

HATEREN, J. H. van; RUDERMAN, D. L. Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, The Royal Society, v. 265, n. 1412, p. 2315–2320, 1998. Citado na página 10.

HOCHBERG, J. E. Effects of the gestalt revolution: The cornell symposium on perception. *Psychological Review*, American Psychological Association, v. 64, n. 2, p. 73, 1957. Citado na página 16.

HUANG, Z. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data mining and knowledge discovery*, Springer, v. 2, n. 3, p. 283–304, 1998. Citado 2 vezes nas páginas 11 e 13.

HUNG, C.-C.; PENG, W.-C. A regression-based approach for mining user movement patterns from random sample data. *Data & Knowledge Engineering*, Elsevier, v. 70, n. 1, p. 1–20, 2011. Citado na página 27.

International Organization for Migration. *International Organization for Migration*. 2015. Disponível em: <https://www.iom.int/world-migration-report-2015>. Citado na página 1.

JIANG, S.; FERREIRA, J.; GONZÁLEZ, M. C. Clustering daily patterns of human activities in the city. *Data Mining and Knowledge Discovery*, Springer, v. 25, n. 3, p. 478–510, 2012. Citado 4 vezes nas páginas x, 3, 30 e 31.

Jiang, S.; Ferreira, J.; Gonzalez, M. C. Activity-based human mobility patterns inferred from mobile phone data: A case study of singapore. *IEEE Transactions on Big Data*, v. 3, n. 2, p. 208–219, June 2017. Citado 3 vezes nas páginas x, 32 e 33.

JMP. *UCommunity Jmp*. 2021. Disponível em: <https://community.jmp.com/t5/JMP-Scripts/Star-Plot-Script/ta-p/21371>. Citado 2 vezes nas páginas ix e 19.

JURDAK, R. et al. Understanding human mobility from twitter. *PLOS ONE*, Public Library of Science, v. 10, n. 7, p. 1–16, 07 2015. Disponível em: <https://doi.org/10.1371/journal.pone.0131469>. Citado na página 33.

KAISER, D. Stick-figure realism: Conventions, reification, and the persistence of feynman diagrams, 1948-1964. *Representations*, University of California Press, v. 70, p. 49–86, 2000. Citado 2 vezes nas páginas ix e 19.

KEIM, D. A. Designing pixel-oriented visualization techniques: Theory and applications. *IEEE Transactions on visualization and computer graphics*, IEEE, v. 6, n. 1, p. 59–78, 2000. Citado 3 vezes nas páginas ix, 17 e 18.

KEIM, D. A. Information visualization and visual data mining. *IEEE transactions on Visualization and Computer Graphics*, IEEE, v. 8, n. 1, p. 1–8, 2002. Citado 6 vezes nas páginas ix, 16, 17, 19, 20 e 21.

KELLER, J. M.; GRAY, M. R.; GIVENS, J. A. A fuzzy k-nearest neighbor algorithm. *IEEE transactions on systems, man, and cybernetics*, IEEE, n. 4, p. 580–585, 1985. Citado 2 vezes nas páginas 8 e 9.

KRISHNA, K.; MURTY, N. M. Genetic k-means algorithm. *IEEE Transactions on Systems Man And Cybernetics-Part B: Cybernetics*, IEEE, v. 29, n. 3, p. 433–439, 1999. Citado 2 vezes nas páginas 9 e 52.

KURITA, T. An efficient agglomerative clustering algorithm using a heap. *Pattern Recognition*, Elsevier, v. 24, n. 3, p. 205–209, 1991. Citado 2 vezes nas páginas 9 e 52.

LIU, S. et al. Vait: A visual analytics system for metropolitan transportation. *IEEE Transactions on Intelligent Transportation Systems*, IEEE, v. 14, n. 4, p. 1586–1596, 2013. Citado na página 35.

LUO, C.; CHUNG, S. M. Efficient mining of maximal sequential patterns using multiple samples. In: SIAM. *Proceedings of the 2005 SIAM International Conference on Data Mining*. [S.l.], 2005. p. 415–426. Citado na página 10.

MACQUEEN, J. et al. Some methods for classification and analysis of multivariate observations. In: OAKLAND, CA, USA. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. [S.l.], 1967. v. 1, n. 14, p. 281–297. Citado na página 11.

Marjani, M. et al. Big iot data analytics: Architecture, opportunities, and open research challenges. *IEEE Access*, v. 5, p. 5247–5261, 2017. Citado 8 vezes nas páginas ix, xiv, 2, 6, 8, 9, 10 e 11.

MUKHOPADHYAY, A. et al. A survey of multiobjective evolutionary algorithms for data mining: Part i. *IEEE Transactions on Evolutionary Computation*, IEEE, v. 18, n. 1, p. 4–19, 2013. Citado na página 8.

NA, S.; XUMIN, L.; YONG, G. Research on k-means clustering algorithm: An improved k-means clustering algorithm. In: IEEE. *2010 Third International Symposium on intelligent information technology and security informatics*. [S.l.], 2010. p. 63–67. Citado na página 12.

NOCKE, T.; SCHLECHTWEG, S.; SCHUMANN, H. Icon-based visualization using mosaic metaphors. In: IEEE. *Ninth International Conference on Information Visualisation (IV'05)*. [S.l.], 2005. p. 103–109. Citado 2 vezes nas páginas ix e 19.

Noulas, A.; Mascolo, C.; Frias-Martinez, E. Exploiting foursquare and cellular data to infer user activity in urban environments. In: *2013 IEEE 14th International Conference on Mobile Data Management*. [S.l.: s.n.], 2013. v. 1, p. 167–176. Citado 4 vezes nas páginas x, 28, 29 e 59.

O'CONNOR, Z. Colour, contrast and gestalt theories of perception: The impact in contemporary visual communications design. *Color Research & Application*, Wiley Online Library, v. 40, n. 1, p. 85–92, 2015. Citado 3 vezes nas páginas ix, 15 e 16.

PACK, M. L. et al. Ice–visual analytics for transportation incident datasets. In: IEEE. *2009 IEEE International Conference on Information Reuse & Integration*. [S.l.], 2009. p. 200–205. Citado 3 vezes nas páginas xi, 39 e 40.

PEIZHUANG, W. Pattern recognition with fuzzy objective function algorithms (james c. bezdek). *SIAM Review*, Society for Industrial and Applied Mathematics, v. 25, n. 3, p. 442, 1983. Citado na página 13.

PLANALTO. *L12587*. 2019. Disponível em: <http://www.planalto.gov.br>. Citado na página 2.

POCO, J. et al. Exploring traffic dynamics in urban environments using vector-valued functions. In: WILEY ONLINE LIBRARY. *Computer Graphics Forum*. [S.l.], 2015. v. 34, n. 3, p. 161–170. Citado 3 vezes nas páginas x, 35 e 36.

PU, J. et al. T-watcher: A new visual analytic system for effective traffic surveillance. In: IEEE. *2013 IEEE 14th International Conference on Mobile Data Management*. [S.l.], 2013. v. 1, p. 127–136. Citado na página 35.

RAVENSTEIN, E. G. The laws of migration. *Journal of the statistical society of London*, JSTOR, v. 48, n. 2, p. 167–235, 1885. Citado na página 1.

ROKACH, L.; MAIMON, O. Clustering methods. In: *Data mining and knowledge discovery handbook*. [S.l.]: Springer, 2005. p. 321–352. Citado 2 vezes nas páginas 14 e 15.

SAGL, G.; LOIDL, M.; BEINAT, E. A visual analytics approach for extracting spatio-temporal urban mobility information from mobile network traffic. *ISPRS International Journal of Geo-Information*, Multidisciplinary Digital Publishing Institute, v. 1, n. 3, p. 256–271, 2012. Citado 4 vezes nas páginas x, 23, 36 e 37.

SCHUBERT, E. et al. Dbscan revisited, revisited: why and how you should (still) use dbscan. *ACM Transactions on Database Systems (TODS)*, ACM New York, NY, USA, v. 42, n. 3, p. 1–21, 2017. Citado na página 9.

SENADO. *Histórico de atualizações da atividade legislativa*. 1988. Disponível em: <https://www.senado.leg.br/atividade/const/con1988/con1988_03.07.2019/art_21_.asp>. Citado na página 2.

SIBSON, R. Slink: an optimally efficient algorithm for the single-link cluster method. *The computer journal*, Oxford University Press, v. 16, n. 1, p. 30–34, 1973. Citado na página 15.

SILVA, E. *Meio ambiente & mobilidade urbana*. Editora Senac São Paulo, 2014. (Série Meio ambiente). ISBN 9788539607341. Disponível em: <https://books.google.com.br/books?id=E_YdogEACAAJ>. Citado na página 2.

SOBRAL, T.; GALVãO, T.; BORGES, J. Visualization of urban mobility data from intelligent transportation systems. *Sensors*, v. 19, n. 2, 2019. ISSN 1424-8220. Disponível em: <https://www.mdpi.com/1424-8220/19/2/332>. Citado 3 vezes nas páginas 16, 23 e 34.

SONG, Y.; MILLER, H. J. Exploring traffic flow databases using space-time plots and data cubes. *Transportation*, Springer, v. 39, n. 2, p. 215–234, 2012. Citado 3 vezes nas páginas x, 3 e 35.

SRIVASTAVA, K. et al. Data mining using hierarchical agglomerative clustering algorithm in distributed cloud computing environment. *International Journal of Computer Theory and Engineering*, IACSIT Press, v. 5, n. 3, p. 520, 2013. Citado na página 9.

SUYKENS, J. A.; VANDEWALLE, J. Least squares support vector machine classifiers. *Neural processing letters*, Springer, v. 9, n. 3, p. 293–300, 1999. Citado 2 vezes nas páginas 8 e 9.

TRAN, P. V.; LE, T. X. Approaching human vision perception to designing visual graph in data visualization. *Concurrency and Computation: Practice and Experience*, Wiley Online Library, 2020. Citado na página 15.

WANG, H. et al. Transportation mode inference from anonymized and aggregated mobile phone call detail records. In: IEEE. *13th International IEEE Conference on Intelligent Transportation Systems*. [S.l.], 2010. p. 318–323. Citado 3 vezes nas páginas x, 27 e 28.

WARD, M. O. Xmdvtool: Integrating multiple methods for visualizing multivariate data. In: IEEE. *Proceedings Visualization'94*. [S.l.], 1994. p. 326–333. Citado 2 vezes nas páginas ix e 19.

World Atlas. *Countries By Number of Cities Over One Million*. 2018. Disponível em: <https://www.worldatlas.com/articles/countries-by-number-of-cities-over-one-million.html>. Citado na página 1.

World Health Organization. *World Health Organization*. 2016. Disponível em: <http://apps.who.int/gho/data/node.main.nURBPOP?lang=enr>. Citado na página 1.

XIONG, H. et al. Mpaas: Mobility prediction as a service in telecom cloud. *Information Systems Frontiers*, Springer, v. 16, n. 1, p. 59–75, 2014. Citado na página 27.

YANG, Z.; KITSUREGAWA, M. Lapin-spam: An improved algorithm for mining sequential pattern. In: IEEE. *21st International Conference on Data Engineering Workshops (ICDEW'05)*. [S.l.], 2005. p. 1222–1222. Citado na página 10.

ZHOU, Y.; REN, Q. Fuzzy c-means clustering algorithm for performance improvement of enn. *Cluster Computing*, Springer, v. 22, n. 5, p. 11163–11174, 2019. Citado na página 13.