



UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

Metodologia para a Obtenção do Nível de Influência de Eventos sobre
Comunidades usando Dados de Mobilidade

Marcos Aurélio de Paiva Souza

Orientadores

Sidney Cunha de Lucena

Carlos Alberto Vieira Campos

RIO DE JANEIRO, RJ - BRASIL

MARÇO de 2021

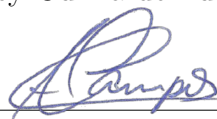
Metodologia para a Obtenção do Nível de Influência de Eventos sobre
Comunidades usando Dados de Mobilidade

Marcos Aurélio de Paiva Souza

DISSERTAÇÃO APRESENTADA COMO REQUISITO PARCIAL PARA OBTENÇÃO DO TÍTULO DE MESTRE PELO PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA DA UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO (UNIRIO). APROVADA PELA COMISSÃO EXAMINADORA ABAIXO ASSINADA.

Aprovada por:

Sidney Cunha de Lucena, DSc. - UNIRIO



Carlos Alberto Vieira Campos, DSc. - UNIRIO

Daniel Sadoc Menasche, DSc. - UFRJ

Ana Paula Couto da Silva, DSc. - UFMG

Ana Cristina Bicharra Garcia, DSc. - UNIRIO

Jefferson Elbert Simões, DSc. - UNIRIO

RIO DE JANEIRO, RJ - BRASIL

MARÇO de 2021

Souza, Marcos Aurélio de Paiva.
Metodologia para a Obtenção do Nível de Influência de Eventos
sobre Comunidades usando Dados de Mobilidade /
Marcos Aurélio de Paiva Souza, 2021.

Orientador: Sidney Cunha de Lucena
Co-Orientador: Carlos Alberto Vieira Campos
Dissertação (Mestrado em Informática) - Universidade Federal do
Estado do Rio de Janeiro, Rio de Janeiro, 2021.

Dedico à minha família, esposa e amigos.

Agradecimentos

Agradeço a minha família que me apoiou em todos os momentos da minha vida.

Agradeço aos meus pais que ajudaram na minha formação como ser humano e me ensinaram a importância da educação.

Agradeço aos meus orientadores Prof. Dr. Sidney Lucena e Prof. Dr. Carlos Alberto Campos por todas as horas de dedicação e paciência ao longo destes anos de trabalho e pesquisa.

Por fim, agradeço a minha futura esposa por toda paciência, incentivo e companheirismo ao longo de todo o período do mestrado. Sem ela eu não chegaria até aqui.

Souza, Marcos Aurélio de Paiva. **Metodologia para a Obtenção do Nível de Influência de Eventos sobre Comunidades usando Dados de Mobilidade**. UNIRIO, 2021. xxx páginas. Dissertação de Mestrado. Departamento de Informática Aplicada, UNIRIO.

RESUMO

Na sociedade moderna, o conhecimento de como eventos influenciam as pessoas e as comunidades é um ponto de extrema importância para o desenvolvimento de políticas públicas e o planejamento de ações orientadas a grupos específicos que possuem uma característica chave em comum. Essas comunidades, formadas por indivíduos que compartilham interesses, comportamentos e semelhanças, possuem características e padrões de mobilidade frequentemente similares devido aos aspectos de comportamento que compartilham.

Dada a evolução tecnológica e a disponibilidade quase ubíqua de sensores GPS nos mais diversos dispositivos, a medida de influência de um evento definido por seu local e período de ocorrência sobre uma comunidade pode ser inferida utilizando dados de mobilidade dos indivíduos dessa comunidade. Assim, este trabalho propõe uma metodologia para a inferência da influência de eventos sobre comunidades utilizando dados de mobilidade dos indivíduos destas comunidades. Esses cálculos se baseiam em informações obtidas do contato de cada indivíduo de uma comunidade com a área de influência de um evento ao longo do tempo de ocorrência deste evento.

De forma a esclarecer esses conceitos e demonstrar suas aplicações em problemas reais, são apresentados três casos de estudos utilizando dados reais de mobilidade urbana na resolução de problemas de otimização aplicados a uma campanha de vacinação e a uma campanha publicitária. Além disso, é apresentada uma comparação utilizando diferentes métodos de criação de comunidades.

Palavras-chave: Análise e Mineração de Dados; Traços de mobilidade; Mobilidade Urbana.

ABSTRACT

In modern society, the knowledge of how events differently influence people and communities is extremely relevant for the development of public policies and the planning of actions aimed at specific groups of citizens that have key features in common. These communities, formed by individuals that share interests, behaviors and similarities, have features and mobility patterns that are often similar due to the behavioral aspects they share.

Given the technological evolution and the almost ubiquitous availability of GPS sensors on many different devices, the measure of influence of an event, defined by its location and period of occurrence, over a community can be inferred using mobility data from the individuals of that community. Thus, this work proposes a methodology to infer the influence of events over communities using mobility data of the individuals of these communities. These calculations are based on information obtained from the contact of each individual of a community with the area of influence of an event over the time of occurrence of this event.

To better understand these concepts and show their applications to real problems, three case studies were conducted using real urban mobility data to solve optimization problems applied to a vaccination campaign and a marketing campaign. Furthermore, a comparison is presented using different methods of creating communities.

Keywords: Data Mining and Analysis; Mobility Trace; Urban Mobility.

Sumário

1	Introdução	1
1.1	Motivação	3
1.2	Objetivos	3
1.3	Contribuições	4
1.4	Organização do Trabalho	5
2	Fundamentação Teórica	6
2.1	Métodos de agrupamento de usuários em comunidades	6
2.1.1	Agrupamento de usuários em comunidades através de seleção manual de atributos	7
2.1.2	Agrupamento de usuários em comunidades através de algoritmos de clusterização	7
2.1.2.1	Algoritmos de clusterização hierárquicos	7
2.1.2.2	Algoritmos de clusterização particionais (Não Hierárquicos)	8
2.1.2.3	Algoritmos de clusterização baseados em distância	8
2.1.2.4	Algoritmos de clusterização baseados em densidade	9
2.1.2.5	Algoritmos de clusterização baseados em modelo (Model based clustering methods [1])	10

2.2	Trabalhos Relacionados	10
3	Modelagem do Problema	16
3.1	Problema	16
3.2	Definições	16
3.2.1	Atributos (A)	16
3.2.2	Usuários (U)	17
3.2.3	Eventos (E)	17
3.2.4	Comunidades (C)	17
3.2.5	Influência	19
3.3	Medidas de interesse	19
3.3.1	Influência Média (I)	19
3.3.2	Influência Relativa (Irel)	19
3.3.3	Influência Média (I) <i>versus</i> Influência Relativa (Irel)	19
3.3.4	Fator de Influência (IF)	20
3.3.5	Matriz de Influência (IM)	20
4	Proposta Metodológica	22
4.1	<i>Workflow</i>	23
4.1.1	<i>Input</i>	24
4.1.2	Pré-Processamento	24
4.1.2.1	Seleção	24
4.1.2.2	Remoção de <i>outliers</i>	25
4.1.3	Processamento dos Dados	25
4.1.3.1	Seleção dos locais de evento	25
4.1.3.2	Seleção da área de influência de eventos	26

4.1.3.3	Seleção do período e duração de eventos	26
4.1.4	Transformação de Dados	27
4.1.4.1	Definição de comunidades	27
4.1.4.2	Criação de comunidades	27
4.1.4.3	Avaliação das comunidades criadas	28
4.1.5	Mineração de dados e cálculo de métricas	28
4.1.5.1	Cálculo de influência média sobre usuários	28
4.1.5.2	Cálculo de influência relativa sobre usuários	28
4.1.5.3	Cálculo de influência média sobre comunidades	29
4.1.5.4	Cálculo de influência relativa sobre comunidades	29
4.1.6	Reconhecimento e Avaliação dos Resultados	30
4.1.6.1	Consolidação dos Resultados	30
4.1.6.2	Visualização dos Resultados	30
4.1.7	<i>Output</i>	30
4.2	Exemplo Aplicado	31
5	Experimentação	41
5.1	<i>Datasets</i>	41
5.1.1	Geolife	41
5.1.2	Cabspotting	44
5.2	Experimento 1	46
5.2.1	Classificação dos usuários em comunidades	47
5.2.2	Definição de Eventos	49
5.2.3	Apresentação dos Resultados	52
5.2.4	Aplicação dos Resultados de Influência - Estudo de Caso 1	56

5.3	Experimento 2	58
5.3.1	Análise e Tratamento dos Dados	58
5.3.2	Classificação dos usuários em comunidades	60
5.3.3	Definição de eventos	61
5.3.4	Apresentação dos Resultados	62
5.3.5	Aplicação dos Resultados de Influência	68
5.3.5.1	Estudo de Caso 1 - Campanha de Vacinação	68
5.3.5.2	Estudo de Caso 2 - Campanha de <i>Marketing</i>	71
5.4	Experimento 3	74
5.4.1	Análise e Tratamento dos Dados	74
5.4.2	Classificação dos usuários em comunidades	75
5.4.2.1	Classificação dos usuários em comunidades utilizando K-Means	76
5.4.2.2	Classificação dos usuários em comunidades utilizando DBSCAN	77
5.4.3	Definição de eventos	80
5.4.4	Apresentação dos Resultados	82
5.4.4.1	Apresentação dos Resultados - K-Means	83
5.4.4.2	Apresentação dos Resultados - DBSCAN	84
6	Conclusão e Trabalhos Futuros	89

Lista de Figuras

3.1	Diagram de Venn com a representação da configuração de comunidades.	18
3.2	Representação da estrutura da Matriz de Influência.	21
4.1	Fluxo de atividade para o cálculo da influência de eventos sobre usuá- rios e comunidades.	23
4.2	Representação do Fator de Influência para o evento e_1 (IF_{e_1}) (Eixo Y) em função do tempo (Eixo X).	32
4.3	Representação do Fator de Influência para o evento e_2 (IF_{e_2}) (Eixo Y) em função do tempo (Eixo X).	33
4.4	Influência Média e Fator de Influência (Eixo Y) do evento e_1 em função do tempo (Eixo X) sobre os usuários u_1 (a), u_2 (b) e u_3 (c). .	33
4.5	Influência e Fator de Influência Média (Eixo Y) do evento e_2 em função do tempo (Eixo X) sobre os usuários u_1 (a), u_2 (b) e u_3 (c). .	34
4.6	Influência Relativa e Fator de Influência (Eixo Y) em função do tempo (Eixo X) para o evento e_1 sobre os usuários u_1 (a), u_2 (b) e u_3 (c). .	35
4.7	Influência Relativa e Fator de Influência (Eixo Y) em função do tempo (Eixo X) para o evento e_2 sobre os usuários u_1 (a), u_2 (b) e u_3 (c). .	36
4.8	Influência Média e Fator de Influência (Eixo Y) em função do tempo (Eixo X) para o evento e_1 sobre as comunidades c_1 (a), c_2 (b) e c_3 (c). .	37
4.9	Influência e Fator de Influência (Eixo Y) em função do tempo (Eixo X) para o evento e_2 sobre as comunidades c_1 (a), c_2 (b) e c_3 (c). . . .	38

4.10	Influência Relativa e Fator de Influência (Eixo Y) em função do tempo (Eixo X) para o evento e_1 sobre as comunidades c_1 (a), c_2 (b) e c_3 (c).	39
4.11	Influência Relativa e Fator de Influência (Eixo Y) em função do tempo (Eixo X) para o evento e_2 sobre as comunidades c_1 (a), c_2 (b) e c_3 (c).	40
5.1	Mapa de calor dos registros presentes no <i>dataset</i> sobre mapa mundi. .	42
5.2	Mapa de calor dos registros sobre o mapa da região da grande Pequim.	42
5.3	Distribuição de registros ao longo do tempo.	43
5.4	Distribuição de registros por dia da semana.	43
5.5	Distribuição de usuários ativos por dia da semana.	44
5.6	Distribuição espacial dos registros utilizando mapa de calor sobre o mapa da cidade de São Francisco (EUA).	45
5.7	Distribuição temporal dos registros por dia da semana.	45
5.8	Distribuição temporal dos registros por data.	46
5.9	Distribuição temporal dos registros por horário.	46
5.10	<i>Clusters</i> dos locais de residência das comunidades para K igual a 5, 11 e 15.	49
5.11	<i>Silhouette score</i> em função de K (a) e Distribuição espacial do local de residência das comunidades para K ótimo ($K = 7$) (b).	50
5.12	Localização geográfica dos locais de ocorrência dos eventos mostrados no <i>Heat Map</i> dos dados no mapa de Pequim.	51
5.13	Resultados obtidos sobre a influência (Eixo Y) para os onze locais de eventos (Eixo X) e para as oito comunidades nas segundas-feiras. . .	52
5.14	Resultados obtidos sobre a influência (Eixo Y) para os onze locais de eventos (Eixo X) e para as oito comunidades nas terças-feiras. . . .	53
5.15	Resultados obtidos sobre a influência (Eixo Y) para os onze locais de eventos (Eixo X) e para as oito comunidades nas quartas-feiras. . . .	53
5.16	Resultados obtidos sobre a influência (Eixo Y) para os onze locais de eventos (Eixo X) e para as oito comunidades nas quintas-feiras. . . .	54

5.17	Resultados obtidos sobre a influência (Eixo Y) para os onze locais de eventos (Eixo X) e para as oito comunidades nas sextas-feiras.	54
5.18	Resultados obtidos sobre a influência (Eixo Y) para os onze locais de eventos (Eixo X) e para as oito comunidades aos sábados.	55
5.19	Resultados obtidos sobre a influência (Eixo Y) para os onze locais de eventos (Eixo X) e para as oito comunidades aos domingos.	55
5.20	Ilustração do processo de remoção de registros duplicados dentro de um intervalo de tempo.	59
5.21	Influência para cada instante de tempo de monitoramento do evento 'E11 - SAT' para os usuários 128 (a), 140 (b), 14 (c) e 155 (d), respectivamente.	63
5.22	Influência para cada instante de tempo de monitoramento do evento 'E11 - SUN' para os usuários 128 (a), 140 (b), 14 (c) e 155 (d), respectivamente.	64
5.23	Influência Média para o evento 'E11 - SAT' para todos usuários de ambas as comunidades (a), e Influência Média do evento 'E11 - SUN' para todos usuários de ambas as comunidades (b).	65
5.24	Influência média do evento "E1" (a), "E2" (b), "E3" (c), "E4" (d), "E5" (e), "E6" (f), "E7" (g), "E8" (h), "E9" (i), "E10" (j) e "E11" (k) em cada dia da semana para todos comunidades. . .	66
5.25	Influência média de todos onze eventos na Segunda (a), na Terça (b), na Quarta (c), na Quinta (d), na Sexta (e), no Sábado (f) e no Domingo (g).	67
5.26	<i>Scatter plot</i> para os locais de início de corrida (a), para os locais de fim de corrida (b) e para os ambos locais de início e de fim de corrida de forma consolidada (c).	75
5.27	Valores obtidos através dos método Elbow (Eixo Y) para cada valor de "k" (Eixo X).	77
5.28	Distribuição espacial dos registros classificados em comunidades utilizando o K-Means.	78

5.29	Valores obtidos após a execução do KNN para o ranqueamento entre as distâncias entre os registros.	79
5.30	Distribuição espacial dos registros classificados em comunidades utilizando o DBSCAN.	79
5.31	Locais de eventos selecionados apresentados sobre o mapa da cidade de São Francisco.	81
5.32	Fator de Influência IF (Eixo Y) em função do tempo (Eixo X).	82
5.33	Influência média (K-Means) do evento “E1” (a), “E2” (b), “E3” (c), “E4” (d), “E5” (e), “E6” (f), “E7” (g), “E8” (h), “E9” (i) e “E10” (j) em cada dia da semana para todos comunidades.	85
5.34	Influência média (K-Means) de todos dez eventos na Segunda (a), na Terça (b), na Quarta (c), na Quinta (d), na Sexta (e), no Sábado (f) e no Domingo (g).	86
5.35	Influência média (DBSCAN) do evento “E1” (a), “E2” (b), “E3” (c), “E4” (d), “E5” (e), “E6” (f), “E7” (g), “E8” (h), “E9” (i) e “E10” (j) em cada dia da semana para todos comunidades.	87
5.36	Influência média (DBSCAN) de todos dez eventos na Segunda (a), na Terça (b), na Quarta (c), na Quinta (d), na Sexta (e), no Sábado (f) e no Domingo (g).	88

Lista de Tabelas

4.1	Horário de entrada e saída da área de influência dos eventos e_1 e e_2 pelos usuários u_1 , u_2 e u_3	32
5.1	Número de usuários por comunidade.	49
5.2	Influência e Custos das ações por dia da semana.	57
5.3	Distribuição ótima de horas por dia da semana obtida através do modelo de otimização proposto.	58
5.4	Lista dos usuários mais influenciados (à esquerda) e lista dos eventos que exerceram influência com maior intensidade (à direita).	62
5.5	Valores de Influência obtidos para cada dia da semana para todos os locais de eventos para a comunidade “C2” para a campanha de vacinação modelada neste estudo de caso.	70
5.6	Distribuição ótima de horas de campanha vacinação por local de evento e por dia da semana com restrição orçamentária.	70
5.7	Valores de Influência Média obtidos para cada dia da semana para todos os locais de eventos para a comunidade C1 para a campanha publicitária.	72
5.8	Distribuição ótima de horas da campanha de marketing por evento e por dia da semana dado uma restrição de influência de evento (alcance) na comunidade C1.	73
5.9	Relação de comunidades identificadas pelo K-Means e a sua quantidade de usuários.	77

5.10 Relação de comunidades identificadas utilizando o DBSCAN e suas respectivas quantidades de usuários.	80
--	----

Lista de Nomenclaturas

ANN	Artificial Neural Networks
BIC	Bayesian Information Criteria
BIRCH	Balanced Iterative Reducing and Clustering Using Hierarchies
CURE	Clustering Using Representatives
DBSCAN	Density-Based Spatial Clustering and Application with Noise
KNN	K Nearest Neighbor
KDD	Knowledge and Discovery on Databases
NN	Neural Network
ROCK	Robust Clustering using links
SMOTE	Synthetic Minority Over-Sampling Technique
SVM	Support Vector Machine
TTR	Travel Time Reliability
UTC	Universal Time Coordinated
V2V	Vehicle to Vehicle

1. Introdução

Nas cidades e metrópoles atuais, eventos, de modo geral, influenciam pessoas e grupos bem como seus comportamentos. Eventos são fatos, ações ou acontecimentos relevantes que produzem efeitos em um ou mais usuários ou grupos de usuários. Assim, esses eventos ocorrem em um determinado local e influenciam usuários e grupos de usuários em uma determinada área durante um certo período de tempo. A intensidade de influência pode ocorrer de maneira variável ao longo do tempo de duração deste evento. Ao falar em eventos, deve-se ter em mente eventos que ocorrem em locais geográficos reais. A título de exemplos de eventos, destaca-se: visualização de propaganda por usuários, acidentes de trânsito, Protestos populares, jogos esportivos, etc.

Comunidades são grupos de pessoas que compartilham entre si características em comum como locais de moradia, padrões de mobilidade, meios de transporte preferidos, etc. Estas comunidades assim como os usuários os quais as pertencem, compartilham e interagem diversos locais geográficos e sofrem os efeitos dos mesmos eventos durante seu cotidiano, podendo perceber os efeitos destes eventos de formas distintas. A título de exemplos de comunidades, podemos ter torcedores de um clube, moradores de um bairro, usuário de um modal de transporte, etc. Segundo [2], comunidades humanas são geralmente formadas com base em: (a) localização, o que inclui aspectos de mobilidade ao longo do tempo, e (b) relacionamentos, como problemas e interesses compartilhados. Além disso, cada indivíduo pode pertencer a uma ou mais comunidades e essas comunidades a que este indivíduo pertence podem e provavelmente serão modificadas ao longo de sua vida.

Assim, dependendo da semântica utilizada para realizar o agrupamento de pessoas em comunidades, é esperado que estas comunidades apresentem diferentes padrões de mobilidade e que indivíduos de comunidades diferentes possuam comportamentos diferentes e estejam expostos de formas diferentes a um evento que ocorre

em um local específico durante um período de tempo. Um exemplo extremamente simples, porém muito esclarecedor, é o caso de uma comunidade formada por trabalhadores rurais. As pessoas desta comunidade frequentam a cidade (região urbana) com pouca frequência e, portanto, a ocorrência de eventos no centro da cidade durante o expediente de trabalho afetará pouco ou nada esta comunidade. De forma similar, o inverso também é verdadeiro. Então, a medida do nível de influência exercida por eventos sobre uma comunidade durante um período de tempo pode ser um recurso extremamente relevante para o planejamento de políticas públicas e para a tomada de decisões relacionadas ao dia-a-dia dessas pessoas e comunidades.

A criação de comunidades pode ser realizada utilizando diversos critérios e métodos, entre os quais figuram métodos que utilizam a seleção de características, através da divisão por faixas de valores destes atributos e algoritmos de classificação. Dentre os algoritmos de classificação de usuários, existem diversas abordagens para realizar esta classificação como as abordagens hierárquicas, particionais, por densidade, entre outras. A escolha do método de classificação de usuários em comunidades vai depender dos dados disponíveis e do objetivo do estudo.

Este trabalho traz como proposta uma metodologia para inferir e expressar como uma comunidade será afetada por um evento. Esta inferência é baseada na estimativa de quantos indivíduos desta comunidade irão perceber a influência deste evento. Para alcançar este objetivo, é realizada a modelagem deste problema e realizadas diversas definições de forma a estabelecer claramente os passos necessários para a obtenção da influência de eventos sobre uma comunidade. É apresentado também um fluxo de atividades desenvolvido para servir como um guia para a aplicação adequada da metodologia proposta para a extração, a partir de uma fonte de dados de mobilidade urbana, do nível de influência de um evento sobre uma determinada comunidade. Este nível de influência é definido como a intensidade que um evento afetou uma comunidade ou usuário ao longo do tempo.

A metodologia proposta é parte de um ferramental extremamente importante para fornecer informações relevantes para diversas áreas de interesse, incluindo as áreas de negócios e governamentais, que dependem do conhecimento do comportamento de pessoas e grupos de pessoas para realizar ações de *marketing*, campanhas de saúde, planejamento de transporte ou planejamento urbano. Por exemplo, em ações de *marketing*, o nível de influência pode auxiliar em alterar dinamicamente painéis digitais, dado um conhecimento prévio da distribuição do público alvo ao longo de uma semana. Em termos de planejamento de eventos, esta métrica pode auxiliar organizações de saúde a escolherem os melhores locais para a condução de

campanhas de vacinação para um grupo de risco baseado no padrão de mobilidade deste grupo através da cidade.

1.1 Motivação

Com o aumento da população em centros urbanos nas últimas décadas, surgiram novos desafios em diversas áreas como sistemas de transporte, saúde, publicidade etc. A necessidade de informações e de entendimento sobre como pessoas vivem, se deslocam e interagem entre si é crucial para o desenvolvimento de soluções que visem a melhorar a qualidade de vida em grandes centros urbanos.

Assim, o entendimento de eventos que afetam positivamente e negativamente pessoas e grupos de pessoas que se relacionam entre si e que possuem características em comum é necessário para fornecer insumos para a elaboração de soluções que auxiliem na escolha e na tomada de decisão voltada para a otimização de cenários. À luz disso, tais lacunas, somadas à ausência de proposta semelhante que abranja múltiplos eventos, bem como a ausência de metodologia simples que possibilite a comparação de eventos de categorias diferentes, são fatores que motivam este trabalho.

1.2 Objetivos

Tem-se como objetivo de estudo a formulação de uma metodologia para a análise e cálculo da influência de eventos sobre usuários e comunidades de forma a ser possível adquirir conhecimento a partir de dados de mobilidade urbana, de forma a utilizá-los em problemas reais que possam ser otimizados de alguma maneira.

Este trabalho tem, portanto, como objetivo responder as seguintes questões de pesquisa:

1. É possível mensurar a influência de eventos sobre comunidades de maneira objetiva e de forma simples?
2. É possível utilizar essa influência para realizar comparações entre diferentes eventos de diferentes tipos e categorias, utilizando uma única unidade de medida?
3. É possível relacionar eventos com comunidades utilizando a influência desses

eventos como fator de comparação?

4. É possível distinguir comunidades utilizando como fator de comparação a influência que um evento exerce sobre elas?
5. É possível comparar eventos e comunidades utilizando semântica atrelada a ambos?

1.3 Contribuições

As principais contribuições deste trabalho são:

1. A proposta de uma metodologia baseada em dados de mobilidade para o cálculo da influência que eventos exercem sobre comunidades formadas por usuários que compartilham características em comum;
2. A elaboração de um conjunto de definições acerca de medidas e relações que não estavam claramente conceituadas na literatura, de modo a formar um entendimento mútuo para o desenvolvimento de pesquisas na área de análise de eventos e seus relacionamentos com usuários e comunidades;
3. A aplicação da metodologia proposta em dois *datasets* bastante difundidos na literatura, GEOLIFE [3] e CABSPOTTING [4], de forma a demonstrar a utilização da metodologia a partir de dados reais de mobilidade urbana;
4. A realização de dois estudos de caso utilizando o *dataset* GEOLIFE para demonstrar a aplicação da metodologia proposta em problemas realísticos. Nestes estudos de **caso**, foram desenvolvidos e respondidos dois problemas de otimização, sendo um para maximar o alcance de uma campanha de vacinação hipotética e o outro para minimizar os custos da realização de uma campanha de *marketing*;
5. Uma comparação entre comunidades criadas a partir do *dataset* CABSPOTTING utilizando métodos diferentes para a classificação de usuários em comunidades, demonstrando assim a importância da criação de comunidades com semânticas bem definidas.

1.4 Organização do Trabalho

No Capítulo 2 são apresentados os métodos de classificação de usuários em comunidades e a descrição de cada um deles. Por fim, são apresentados os trabalhos relacionados que fornecem a fundamentação teórica para o desenvolvimento da metodologia proposta.

No Capítulo 3 são apresentados os conceitos de Evento, Comunidade, Influência, Fator de Influência e as características de cada um deles.

No Capítulo 4 é apresentada a proposta deste trabalho e o detalhamento de cada uma das etapas e dos passos do *workflow* definido. Neste capítulo também são formuladas as equações para o cálculo de Influência e é apresentada uma ilustração de sua aplicação.

No Capítulo 5 são apresentadas as fontes de dados utilizadas, demonstrando suas principais características e descrevendo três experimentos em que cada um explora, de forma diferenciada, as possíveis configurações da metodologia proposta, apresentando aplicações em cenários reais com dados hipotéticos.

Por fim, são apresentadas, no Capítulo 6, as conclusões deste trabalho e trabalhos futuros a serem desenvolvidos a partir do arcabouço disponibilizado e estruturado neste trabalho.

2. Fundamentação Teórica

Neste Capítulo são apresentados os conceitos e métodos a respeito do agrupamento de usuários em comunidades. Além disso, são apresentados trabalhos relacionados que foram identificados durante a revisão da literatura e que suportam as conceituações e as propostas desenvolvidas nesta dissertação.

2.1 Métodos de agrupamento de usuários em comunidades

Existem diversos métodos de agrupamento de usuários em comunidades desenvolvidos e disponíveis na literatura. Estes métodos têm como objetivo realizar a separação de usuários em grupos que possuam características em comum, o que possibilita que seja realizada a identificação de cada um destes grupos através de uma semântica clara de comunidade.

Para que o agrupamento de usuários em comunidades ocorra, faz-se necessária a definição dos critérios com base na qual será realizado este agrupamento. Estes critérios devem se basear em atributos dos usuários de forma que cada grupo ou comunidade contenha um significado distinto, ou seja, possua uma semântica que a defina.

Os métodos disponíveis para a realização do agrupamento de usuários em comunidades podem ser classificados em relação à forma que o agrupamento é realizado. A forma que este agrupamento ocorre pode ser dividida em dois grupos: agrupamento por seleção manual de atributos ou agrupamento por algoritmos de clusterização.

2.1.1 Agrupamento de usuários em comunidades através de seleção manual de atributos

No agrupamento por seleção manual, é realizado o agrupamento a partir da escolha de um ou mais atributos de forma que estes sejam divididos em faixas de valores, com isso cada faixa de valor representa uma comunidade.

A título de exemplo, cita-se o agrupamento pelo atributo “idade”. Seja o atributo “idade” subdividido em faixas etárias: criança (0 a 12 anos), adolescente (13 a 18 anos), adulto (18 a 59 anos), e idoso (60 anos ou mais). A partir desta divisão, pode-se definir quatro comunidades com a semântica de cada comunidade atrelada a uma faixa etária.

2.1.2 Agrupamento de usuários em comunidades através de algoritmos de clusterização

Neste método de agrupamento de usuários em comunidades, o agrupamento é realizado por algoritmos de clusterização que realizam o agrupamento destes usuários utilizando atributos.

Segundo [5], clusterização é a classificação não supervisionada de padrões (observações, dados, ou características) em grupos chamados de *clusters*. O termo clusterização é utilizado para descrever os métodos de agrupamento de dados não rotulados. Métodos de clusterização são bastante difundidos e existem inúmeros *surveys* disponíveis na literatura sobre o assunto. A título de exemplo, citam-se [5], [6], [7], [1] e [8].

Os algoritmos de clusterização utilizam métricas de medida de distância, similaridade ou dissimilaridade para verificar o quão próximo ou distante um elemento está de outro. Dentre as principais medidas de distância, estão a (a) Euclidiana, a (b) Manhattan (blocos) e a (c) Geodésica.

Segundo [1], existem dois principais tipos de métodos de clusterização: (a) Hierárquicos e (b) Particionais.

2.1.2.1 Algoritmos de clusterização hierárquicos

Os algoritmos de clusterização hierárquicos utilizam normalmente uma abordagem divisiva (*top-down*) ou aglomerativa (*bottom-up*) para a classificação de dados em *clusters*.

Na abordagem divisiva, cria-se inicialmente um único cluster com todos usuários disponíveis. Em seguida, é realizada a divisão em dois *clusters* menores com a maior distância *inter-cluster*. Após a divisão, repete-se a divisão de cada *cluster* até que todos *clusters* possuam apenas um elemento.

Já na abordagem aglomerativa, cria-se inicialmente um número de *clusters* tão grande quanto forem o número de elementos disponíveis. A partir disso, realiza-se a junção de *clusters* que possuam a menor distância *inter-cluster*. Após a junção, repete-se a junção dos *clusters* até que todos *clusters* tenham formado um único *cluster*.

O método hierárquico estabelece uma hierarquia entre os *clusters* formados a medida que vai formando estes *clusters*. Este método não necessita de uma definição a priori do número “k” de clusters, entretanto, devido a sua complexidade computacional, sua utilização não é aconselhada em cenários com um número elevado de dados a serem processados. Alguns exemplos de algoritmos de clusterização hierárquicos são: ROCK (Robust Clustering using links) [9], BIRCH (Balanced Iterative Reducing and Clustering Using Hierarchies) [10] e CURE (Clustering Using Representatives) [11].

2.1.2.2 Algoritmos de clusterização particionais (Não Hierárquicos)

Já os algoritmos de clusterização particionais (não hierárquicos) realizam a classificação de usuários em “k” *clusters*. Estes algoritmos de clusterização são chamados de particionais, pois realizam o particionamento dos dados em grupos não hierarquizados entre si. Dentre os algoritmos de clusterização particionais, existem três principais abordagens de clusterização: (a) Baseada em distância; (b) Baseada em Densidade; e (c) Baseada em modelo.

2.1.2.3 Algoritmos de clusterização baseados em distância

Os algoritmos de clusterização baseados em distância realizam a classificação de usuários calculando a distância entre usuários de forma a minimizar a distância *intra-cluster* e maximizar a distância *inter-cluster*.

Um dos algoritmos mais bem sucedidos e eficientes em termo de tempo de processamento e facilidade de implementação desta abordagem é o K-Means. O K-Means e suas variações são amplamente utilizados na literatura, seja como algoritmos de *baseline* para comparações de eficiência, eficácia e acurácia em casos de proposição de novos algoritmos e métodos de classificação, ou como algoritmo principal para

clusterização, devido a sua excelente **desempenho** em processamento em função de sua baixa complexidade computacional.

O K-Means necessita que um número “k” de *clusters* seja definido a priori para que ele possa ser utilizado. Para que o K-Means ofereça uma boa classificação é necessário que o valor de “k” escolhido seja mais próximo do valor ótimo quanto possível.

Existem diversos métodos para a escolha do valor “k”. A título de exemplo, podemos citar os seguintes métodos e artigos que os utilizaram: (a) Método Elbow [12]; (b) Silhouette Score [12] e [13]; (c) Bayesian Information Criteria (BIC) [14]; e (d) Davies Bouldin [15].

O método Elbow utiliza uma heurística relacionada ao ganho de informação para cada *cluster* “k” analisado. Para isto, é gerada uma figura correlacionando a variância do elementos de cada cluster com o número “k” de *clusters*. A partir da análise deste gráfico, visa-se a escolha de um “k” próximo à região da figura onde o ângulo formado pela reta entre dois pontos (ganho de informação) é pequeno.

Já o método do Silhouette Score calcula o quão similar um ponto é do seu próprio *cluster* quando comparado com todos os demais *clusters*. Como detalhado em [16], o cálculo do *Silhouette Score* é obtido da seguinte forma:

- Para cada elemento j , calcular a distância média desse elemento para todos os demais elementos de seu *cluster*, representado por X_j ;
- Para cada elemento j , calcular a menor distância entre o elemento j e todos os elementos de todos os *clusters*, representado por Y_j ;
- Para cada elemento j , o valor do *Silhouette Score* é dado por $S_j = \frac{(Y_j - X_j)}{\max(X_j, Y_j)}$.

2.1.2.4 Algoritmos de clusterização baseados em densidade

Os algoritmos de clusterização baseados em densidade realizam a classificação dos usuários em comunidades utilizando o critério de densidade de dados. Os *clusters* são formados a partir de uma quantidade mínima de dados definida. Ao contrário dos algoritmos de clusterização baseados em distância, os algoritmos baseados em densidade são menos afetados por *outliers*, uma vez que *outliers* rotineiramente não são densos. Além disso, os *clusters* formados podem possuir formatos não lineares, uma vez que não dependem apenas da distância para a formação desses.

Dentre os algoritmos baseados em densidade, destaca-se o DBSCAN, que assim como o K-Means é um dos algoritmos mais utilizados na literatura. Como demonstrado em [17], o DBSCAN não necessita de um número “k” de *clusters* a priori. Em contrapartida, sua implementação necessita de dois parâmetros: (a) *eps* e (b) *min_samples*. O parâmetro “*eps*” serve para definir a distância máxima entre dois pontos para que ambos sejam considerados como sendo de uma mesma vizinhança. Já o parâmetro “*min_samples*” serve para definir o número mínimo de elementos para que se crie um *cluster*.

2.1.2.5 Algoritmos de clusterização baseados em modelo (Model based clustering methods [1])

Algoritmos de clusterização baseados em modelos utilizam modelos matemáticos, ou a mistura de modelos, de forma a encontrar dados compatíveis e atribuí-los a *clusters* [18]. Os algoritmos mais comuns e utilizados, segundo [1], são as árvores de decisão (decision trees) e as redes neurais (Neural Networks - NN).

Estes algoritmos são utilizados em casos onde a distribuição dos dados ocorre conforme uma distribuição normal ou Guassiana, ou uma mistura de distribuições conforme uma dada função de probabilidade.

2.2 Trabalhos Relacionados

Nesta seção, são expostos e discutidos trabalhos relacionados com a metodologia proposta e que será apresentada no Capítulo 4. Estes trabalhos suportam as ideias e os conceitos aqui apresentados auxiliando no entendimento e desenvolvimento desta metodologia em três principais pilares, sendo eles: (a) Análise de influência de eventos em usuários e comunidade; (b) Métodos de classificação de usuários utilizando algoritmos de clusterização; e (c) Análise de características e de comportamento de usuários e mobilidade urbana.

Em [19] é realizada uma análise do impacto de chuvas no fluxo do tráfego na cidade de Pequim. Os autores correlacionam variáveis como volume de chuvas, velocidade e fluxo de veículos para mensurar o impacto de chuvas no tráfego de veículos. Entre as contribuições apresentadas, as mais relevantes para a nossa metodologia são: (a) o nível de detalhamento em que cada evento e seus efeitos sobre o tráfego de veículos nas vias urbanas da cidade de Pequim; (b) a comprovação de que é possível correlacionar variáveis (atributos) de eventos com usuários; e (c) a análise estatística

dos resultados demonstrando uma relação de causa-efeito entre eventos climáticos e as características do fluxo de veículos em vias urbanas analisadas. Apesar das contribuições significativas expostas por [19], os autores propuseram um estudo sobre uma área muito restrita da cidade de Pequim para realizar a análise do efeitos das chuvas sobre o tráfego de veículos, o que pode gerar vieses introduzidos por características e comportamentos dos usuários daquelas vias ou devido a características das próprias vias que podem ser diferentes de outras vias da cidade.

Mesmo que [19] não tenha utilizado o termo "evento", chuvas e outras ocorrências ocasionadas por condições climáticas como neve, geadas e outros fatores são tratados como eventos pela metodologia proposta. Após ser estabelecido este paralelo entre eventos, é possível utilizar os resultados e modelos criados por [19] como **entrada** para realizar o cálculo da influência de um evento sobre usuários e comunidades.

Em [20] também é analisado o impacto de condições climáticas na mobilidade urbana com um enfoque em enchentes. De forma similar a metodologia desenvolvida por [19], [20] emprega uma análise especializada em apenas um tipo de evento, diferentemente da aqui proposta. Com esse objetivo, [20] define o risco de enchentes utilizando como variáveis: (a) a probabilidade de ocorrência de uma enchente; (b) o impacto de uma enchente na mobilidade urbana. Como contribuição de [20] pode-se destacar a utilização de correlação entre variáveis como condições climáticas e características do local de ocorrência de enchentes com a mobilidade urbana de forma diferente do apresentado em [19].

Em [21] são analisadas as condições meteorológicas e como elas afetam o tráfego em uma região. Para essa análise, é realizado um levantamento do volume de tráfego, número de acidentes, gravidade dos acidentes ocorridos e velocidade média de deslocamento em diversas condições meteorológicas. Como resultado, concluiu-se que as condições meteorológicas de fato impactam na mobilidade urbana, podendo aumentar ou diminuir a capacidade da vias, quantidade e gravidade de acidentes e velocidade. Apesar de ser uma boa contribuição, [21] analisa apenas condições climáticas como fator determinante de impacto na mobilidade. Assim, não são analisadas as características do condutor como idade, gênero ou veículo utilizado, sendo que tais aspectos, segundo trabalho desenvolvido em [22], relacionam-se com grupos mais propensos à participação em acidentes.

Em [23] é investigada a resiliência da mobilidade humana. Os autores a definem como sendo o nível de liberdade que uma pessoa pode se locomover através de uma rede viária dentro de uma determinada área. Foram analisados os dados de um

período de dez anos compreendidos no intervalo dos anos 2009 e 2019. Durante o período analisado, ocorreram grandes eventos climáticos como o Furacão Sandy. Dentre as contribuições deste trabalho, a análise de eventos climáticos raros, chamados pelos autores de eventos extremos, se destaca. A análise de eventos raros, como o furacão Sandy, é extremamente relevante devido a sua grande importância e alto grau de influência sobre diversos usuários e comunidades na área de influência deste evento. Além disso, estes eventos são mais simples de serem analisados devido sua grande influência e área de influência o que minimiza a interferência de outros eventos durante a análise do comportamento do evento. Entretanto, para a realização da análise destes eventos, [23] utilizaram como principal fonte de dados, dados de redes sociais e dados deste tipo podem ser extremamente enviesados devido a fatores como: (a) frequência de postagens; (b) padrões de utilização do usuário; (c) necessidade de infraestrutura disponível para o envio de postagens; (d) entre outros fatores. Estes tipos de eventos, assim como diversos outros tipos de eventos, estão previstos na metodologia proposta.

Em [24] é realizada uma análise voltada para a forma como acidentes de trânsito afetam a confiabilidade do tempo de viagem (TTR) para uma determinada rota. Os autores utilizaram, para esta análise, dados do histórico de tráfego que estavam disponíveis para um grande rodovia. A partir destes dados, realizaram a decomposição em diversas variáveis, como por exemplo, periodicidade e sazonalidade, que poderiam afetar a TTR.

Como importantes contribuições a serem mencionadas, destaca-se a análise e a extração de acidentes de tráfego. Estes são considerados como evento na metodologia proposta, bem como a maneira pela qual eles afetam a TTR. A análise e a extração de eventos podem ser aplicadas na metodologia proposta na etapa de definição e escolha de eventos a serem analisados. Conforme explicado mais adiante na Seção 4.1.3.1, eventos podem ser selecionados por meio de uma análise realizada a partir de dados disponíveis de mobilidade urbana. Outra contribuição de [24] é a forma em que se foi correlacionada a TTR com os diferentes graus de acidentes de trânsito, em outras palavras, como um acidente com o envolvimento de apenas um carro afeta a TTR de forma diferente quando comparado com um acidente envolvendo diversos carros. De forma geral, a análise realizada em [24] auxilia no entendimento sobre maneiras de identificações de alguns tipos de eventos utilizando dados de tráfego.

Como ponto negativo, porém, condições climáticas não foram consideradas, o que pode adicionar erros no modelo de predição proposto por [24]. Outro ponto negativo é a utilização de dados no intervalo de apenas um ano em uma pequena área

da Austrália, o que somado a não utilização de outros *datasets* para a validação do modelo proposto, prejudica a confirmação dos resultados obtidos e a generalização do modelo. Por fim, não são realizados cálculos de quantos usuários e quantas comunidades são efetivamente afetados por estes eventos, sendo realizada a medição estritamente do atraso adicional em uma viagem.

Em [25] é mostrado um novo método de classificação de usuários em comunidades baseadas em similaridade de mobilidade. Neste método, todos os usuários são classificados em comunidades utilizando características como meio de transporte, velocidade de movimento e posicionamento espacial ao longo do tempo. [25] demonstra como é possível realizar o agrupamento destes usuários de maneira eficiente, de forma a otimizar o roteamento de informações exercendo um menor *overhead* em um rede oportunística formada com os usuários. Esse trabalho contribui para a metodologia proposta apresentando uma nova forma de agrupamento de usuários em comunidades. Este novo modo de agrupamento de usuários em comunidades pode ser utilizado na etapa de classificação de usuários em comunidades, descrita em detalhes na seção 4.1.4.2. Assim, são criadas comunidades baseadas em padrões de mobilidade de forma a contribuir ao objetivo da metodologia proposta de avaliar e planejar eventos, maximizando ou minimizando seus efeitos conforme desejado.

Em contraste com a metodologia proposta, [25] não apresenta como escopo a análise da influência de eventos em comunidades, focando apenas na problemática de classificação de usuários em comunidades.

Em [26], os autores desenvolvem a problemática de identificação de comunidades e propõem um novo método de detecção de comunidades sobrepostas chamada *TO-TAR Framework*. O método proposto utiliza as características de usuários e suas preferências ao longo do tempo de maneiras distintas, como a análise de condições climáticas. Como contribuição de [26], frisa-se o método de detecção de comunidades sobrepostas que pode auxiliar na etapa de criação de comunidade que é apresentada na seção 4.1.4.2, gerando sinergias e possibilitando a utilização de novas semânticas para as comunidades criadas. Como ponto negativo de [26], cita-se que o modelo proposto necessita de validação utilizando outras fontes e tipos de dados, sendo imprescindível a verificação de sua aplicabilidade em outros tipos de problemas.

Em [27], é tratada a questão de detecção de comunidades a partir da análise de dados disponíveis em debates públicos/políticos. Para este fim, [27] utilizou uma heurística para gerar grafos com o relacionamento entre usuários e comunidades, realizando a definição de pesos e direcionamento do relacionamento de cada um dos

nós do grafo. Como contribuição deste trabalho um novo método para correlacionar comunidades baseado no alcance de cada debate entre estas comunidades é apresentada. Entretanto, não foi explorado todas as possibilidades que foram expostas no artigo, além de não ter endereçado o problema de relacionamento de um usuário com múltiplas comunidades. Nesta direção, a metodologia proposta no presente trabalho lida com a questão de múltiplos relacionamentos entre usuários e comunidades de forma transparente, permitindo a existência dessa característica nas comunidades criadas.

Em [28], é apresentada uma análise de redes sociais baseadas em eventos, sendo proposto um método de detecção de comunidades baseadas na influência que cada usuário exerce sobre os demais. Para este fim, [28] criou uma função de peso para realizar o balanceamento das duas variáveis presentes no modelo proposto por ele: (a) influência social baseada em estrutura; e (b) influência social baseada em comportamento. Esta função tem como objetivo realizar o balanceamento das relações online e offline entre os usuários. De forma a realizar a classificação dos usuários em comunidades, foi utilizado o algoritmo de clusterização K-Means utilizando como **entrada** a influência social medida. Como contribuição, pode-se mencionar a proposta de realização de clusterização de usuários através do nível de influência de eventos sobre os usuários ao invés de classificá-los utilizando atributos característicos aos usuários.

Por outro lado, ao contrário do apresentado [28], que propõe métodos de detecção de comunidades, a metodologia proposta foca na questão da avaliação e comparação de eventos utilizando como característica de comparação a influência geradas por estes eventos sobre comunidades. Além disso, enquanto [28] realiza uma análise de interações virtuais, o objetivo da nossa metodologia é a análise principalmente de eventos que ocorrem em locais reais possuindo locais físicos de ocorrência, mas não se limitando apenas a estes cenários.

Em [29], é realizada uma análise detalhada sobre as características e os padrões do comportamento dos taxistas de São Francisco (EUA). Características como: (a) perfil da velocidade do taxistas; (b) distribuição espaço temporal dos taxistas; (c) distribuição das frequência de início e fim de corridas; (d) identificação de *hotspots*; (e) duração média de corridas e tempo sem passageiro; (f) conectividade V2V; e (g) particionamento e clusterização de redes.

Apesar de realizar um detalhamento excelente e bastante minucioso do *dataset* de táxis de São Francisco [4], [29] não demonstra efetivamente a identificação e seleção

de *hotspots*. O que é demonstrado é a localização de cada ponto de início e fim de cada corrida plotado sobre o mapa da cidade de São Francisco. Outro ponto negativo é o particionamento e clusterização dos táxis. Em ambos os casos não foi possível identificar de forma clara os passos seguidos em ambos os casos e nem a aplicação pretendida com os resultados obtidos. Apesar dos pontos negativos, [29] contribui com sua análise detalhada de um *dataset* bastante utilizado na literatura, de forma a diminuir o esforço da análise prévia dos dados deste *dataset* tão importante que será utilizado na Seção 5 para a demonstração da metodologia proposta.

Apesar da existência de uma vasta quantidade de trabalhos em sistemas de recomendação [30] [31], em análise de impactos na mobilidade urbana [32], [33], e em análise de trajetórias [34], [35], [36], não foi possível identificar trabalhos que investigassem formação e utilização de comunidades baseadas em múltiplas características voltadas para a análise de padrões e problemas de mobilidade urbana de forma a correlacionar estes padrões com comunidades, usuários e eventos. Além disso, apesar da existência de artigos como [37], [38], [39] e [21], que analisam e desenvolvem pesquisas sobre como eventos afetam usuários, comunidades e a mobilidade urbana, não foi possível identificar uma definição para “evento” como algo que englobasse, de alguma forma, todos os tipos de eventos que podem exercer algum tipo de influência sobre usuários, comunidades e na mobilidade urbana de um local. Cada um destes trabalhos focam em um ou em um grupo pequeno de eventos e, devido a isto, não definem eventos apropriadamente.

A metodologia proposta no presente trabalho, ao contrário dos demais trabalhos encontrados na literatura, desenvolve uma análise acerca de eventos de uma maneira mais genérica e ampla. Nesse sentido, foi necessário definir formalmente o termo evento, de modo a criar um ponto de partida para que se fosse possível iniciar discussões e gerar comparações entre eventos e seus efeitos, bem como sua influência sobre usuários e comunidades.

3. Modelagem do Problema

Neste capítulo, apresentaremos uma modelagem contendo as definições e os conceitos necessários para o entendimento do problema investigado nesta dissertação.

3.1 Problema

A análise de eventos e seus efeitos sobre comunidades é um assunto pouco explorado na literatura. A necessidade de se obter informações sobre o relacionamento entre eventos e comunidades e os efeitos que estes eventos exercem sobre usuários é importante para viabilizar a proposição de meios que otimizem o alcance e a influência destes eventos como desejado

Assim, tem-se como objetivo deste estudo, a criação de um método para o cálculo da influência de eventos sobre usuários e comunidades de forma a ser possível adquirir conhecimento a partir de dados de mobilidade urbana, de forma a utilizá-los em problemas reais que possam ser otimizados de alguma maneira.

3.2 Definições

3.2.1 Atributos (A)

Primeiramente, define-se atributo como qualquer característica a ser utilizada com o intuito de distinguir um grupo de usuários de outro grupo de usuários. Seja $A = \{a_1, \dots, a_N\}$ o conjunto de características existentes. A título de exemplos de atributos, temos: características socioeconômicas, locais de moradia, meio de transporte favorito, locais de lazer, etc. Sendo N a cardinalidade do conjunto de atributos (A).

3.2.2 Usuários (U)

Usuário é definido como qualquer pessoa que teve seu deslocamento monitorado ao longo de um certo período de tempo. Esta pessoa deve estar identificada através de um identificador único de forma que seja possível distingui-la de outro usuário. Assim, cada usuário possui um conjunto de atributos que proporcionam a aquisição de características e de padrões de mobilidade distintas de outros usuários.

Seja $U = \{u_1, \dots, u_O\}$ um conjunto de usuários e A o conjunto de atributos que esses usuários podem possuir, então $A_{u_o} = \{a_1, \dots, a_P\}$ é o subconjunto de atributos do conjunto A que define o usuário u_o . As variáveis O e P referem-se respectivamente ao último elemento do conjunto de usuários (U) e à quantidade de elementos do conjunto de atributos do usuário u_o (A_{u_o}). Já a variável o refere-se à representação de um dado elemento do conjunto de usuários (U).

Como forma de ilustrar, alguns exemplos de usuários podem ser táxis, pedestres, ônibus, pessoas em múltiplos modais de transporte etc.

3.2.3 Eventos (E)

Evento é a ocorrência de qualquer fato ou ação relevante que gera efeito em um ou mais usuários ou grupos de usuários. Cada evento possui características como local em que o evento ocorre, área de influência, frequência de ocorrência, duração, data e hora de início e data e hora de fim. A título de exemplos de eventos, temos: visualização de propaganda por usuários, acidentes de trânsito, protestos populares, jogos esportivos, etc.

Um conjunto de eventos pode ser descrito representado por $E = \{e_1, \dots, e_l, \dots, e_L\}$, onde L indica a quantidade total de eventos e relaciona-se ao último elemento desse conjunto E . A variável l refere-se à representação de um dado elemento do conjunto de eventos E .

3.2.4 Comunidades (C)

Comunidade é definido como um grupo de usuários que compartilham entre si características em comum, ou seja, atributos em comum. Para se definir uma comunidade c_m , é necessário selecionar um subconjunto de atributos A_{c_m} do conjunto A e classificar os usuários em grupos utilizando estes atributos como critério.

Seja $C = \{c_1, \dots, c_M\}$ o conjunto de comunidades e que cada comunidade c_m é definida por um subconjunto de atributos $A_{c_m} = \{a_1, \dots, a_R\}$, então temos que cada comunidade $c_m = \{u_1, \dots, u_Q\}$ é composta por Q usuários, de forma que $\forall u_q, u_q \in c_m \leftrightarrow A_{c_m} \subseteq A_{u_q}$.

As variáveis M , Q e R referem-se ao último elemento do conjunto de comunidades (C), do conjunto de usuários da comunidade c_m e do conjunto de atributos que definem a comunidade c_m , respectivamente. Já as variáveis m e q se referem a representação de um dado elemento do conjunto de comunidades (C) e do conjunto dos usuários da comunidade c_m , respectivamente.

A título de exemplos de comunidades, podemos ter torcedores de um clube, moradores de um bairro, usuário de um modal de transporte, etc.

Com o escopo de ilustrar a relação entre os usuários e as comunidades, observa-se a Figura 3.1. É possível notar o relacionamento entre oito usuários e quatro comunidades, onde cada usuário possui três atributos: gênero (Homem/Mulher), situação empregatícia (Empregado/Não empregado) e situação estudantil (Estudante/Não estudante).

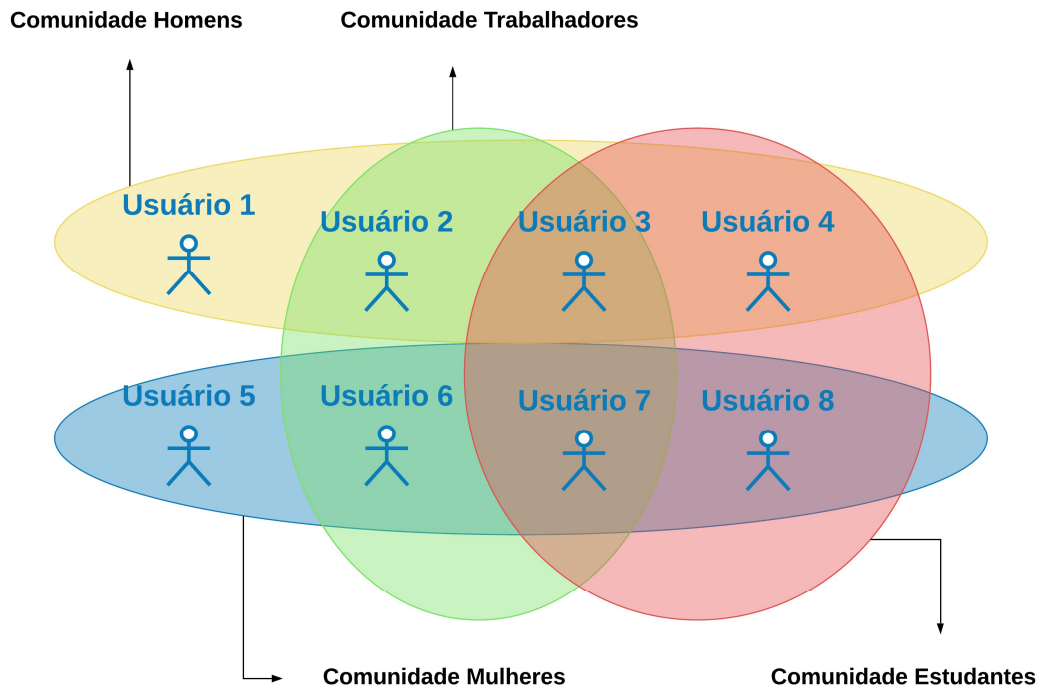


Figura 3.1: Diagram de Venn com a representação da configuração de comunidades.

3.2.5 Influência

Influência é definida como o efeito que um evento exerce sobre usuários e comunidades. Este efeito pode ser um tempo maior de viagem do que o esperado, a obtenção de alguma informação através de anúncios em vias ou *outdoors*, etc. Este efeito é determinado pelo tempo em que cada usuário u_o esteve dentro da área de influência de um determinado evento e_l e pelo Fator de Influência de e_l (IF_{e_l} , a ser definido mais adiante). É dito que um evento e_l influencia um usuário u_o se e somente se u_o estiver dentro da área de influência de e_l durante o período de ocorrência de e_l .

3.3 Medidas de interesse

A partir das definições apresentadas acima, são descritas abaixo as medidas de interesse necessárias para a análise do impacto dos eventos nos usuários e comunidades. As expressões para os cálculos destas medidas serão apresentadas no capítulo seguinte.

3.3.1 Influência Média (I)

Influência Média é a medida de influência que um evento e_l exerce sobre um usuário u_o ao longo de um período de tempo (dia, semana, mês, etc.), mesmo que e_l não ocorra continuamente durante todo este período de tempo.

3.3.2 Influência Relativa (Irel)

Influência Relativa é a influência que um evento e_l exerce sobre um usuário u_o ao longo de um período de tempo (dia, semana, mês, etc.), desconsiderando os intervalos de tempo em que o evento não está ocorrendo.

3.3.3 Influência Média (I) *versus* Influência Relativa (Irel)

A principal diferença entre Influência média e Influência relativa é a forma como cada uma trabalha o aspecto da variação de duração de cada evento.

A Influência Média auxilia em uma análise com mais ênfase na duração do evento. Assim, na hipótese de eventos apresentarem a mesma área de influência

sobre os usuários e as comunidades com as mesmas características, eventos que possuem maior duração tendem a influenciar mais usuários do que eventos que possuem duração menor.

Em contrapartida, a Influência Relativa é melhor para analisar e comparar eventos com durações de tempo diferentes ou consideráveis distinções nos padrões de comunidade e de usuários analisados. Isso se deve ao fato do tempo de observação dos usuários não influenciar diretamente o cálculo de influência relativa.

3.3.4 Fator de Influência (IF)

Fator de Influência é a intensidade com que um evento e_l afeta usuários e comunidades ao longo do tempo de duração de e_l . O valor do Fator de Influência de e_l (IF_{e_l}) pode ser variável ou constante ao longo do período de duração de e_l . A função $IF_{e_l}(t)$ representa a evolução da IF_{e_l} em função do tempo e os valores que a função assume para cada instante t ($IF_{e_l}(t)$) devem ser tais que: $IF_{e_l}(t) \in \mathbb{R}_+$.

Os valores assumidos pela função $IF_{e_l}(t)$ podem ser considerados constantes ao longo do tempo em casos nos quais os eventos são simples ou em casos que $IF_{e_l}(t)$ é desconhecida. Nestes casos, assume-se $IF_{e_l}(t) = 1$.

Nos demais casos, $IF_{e_l}(t)$ assume valores variáveis de acordo com estudos prévios com a análise de causa-efeito do evento sobre usuários e comunidades. Exemplifica-se esse cenário com eventos como os causados por condições climáticas, como enchentes em que a profundidade de água nas ruas dita se a influência é maior (inviabilizando o trânsito por uma via) ou menor (diminuindo a velocidade média dos usuários).

É importante ressaltar que, o Fator de Influência é uma característica intimamente relacionada com o evento em si. Assim, mesmo dentro de um mesmo estudo onde são analisados diversos eventos, os Fatores de Influência de cada um dos eventos são independentes entre si, dependendo exclusivamente do próprio evento. Desta forma, é importante realizar a análise prévia de cada evento de forma a possuir os Fatores de Influência para cada evento antes da realização do estudo.

3.3.5 Matriz de Influência (IM)

Matriz de Influência é uma matriz tridimensional com dimensões $l \times o \times T$ que relaciona cada usuário u_o com cada evento e_l para todo instante de tempo t . Assim, cada coluna de IM representa um evento e_l (eixo x), cada linha de IM representa

um usuário u_o (eixo y) e cada camada representa um instante de tempo t (eixo z).

Cada célula da matriz IM é identificada pela variável IM_{lot} e assume valor “1” se, para aquele instante de tempo t , o usuário u_o está dentro da área de influência do evento e_l ou zero, caso contrário. A Figura 3.2 ilustra a estrutura da Matriz de Influência.

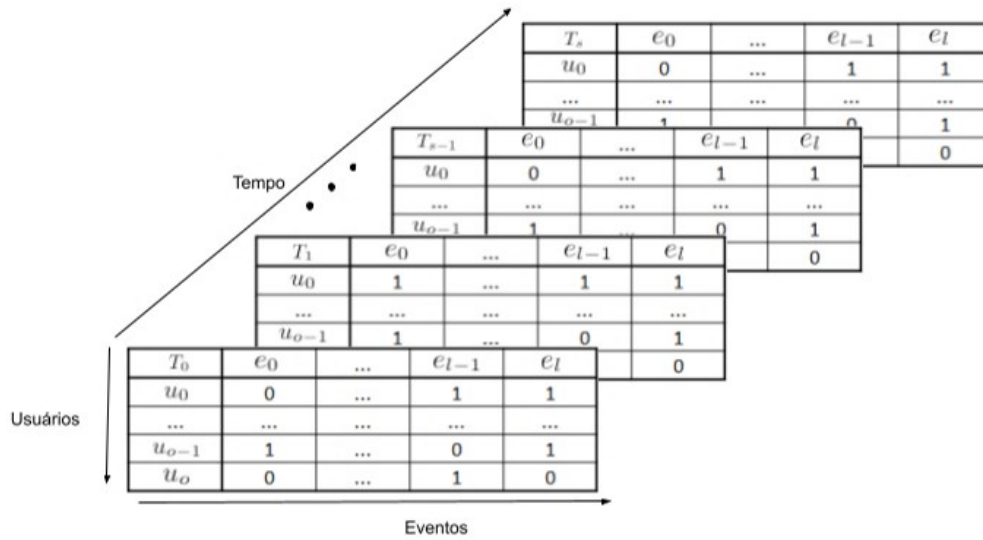


Figura 3.2: Representação da estrutura da Matriz de Influência.

A partir das definições e medidas de interesse descritas neste capítulo, será definida a metodologia proposta com a finalidade de quantificar o nível de influência de eventos sobre comunidades formadas por critérios de localização.

4. Proposta Metodológica

Neste capítulo é apresentada a metodologia proposta, qual seja, a execução dos passos de análise a partir da definição de eventos e comunidades, realizando o tratamento desses dados. Ademais, inserem-se também a execução do cálculo de influência para esses eventos sobre usuários e comunidades e a consolidação dos resultados de maneira a gerar conhecimento aplicável em casos reais. Por fim, são apresentados também os principais motivadores para o desenvolvimento desta metodologia, possíveis cenários favoráveis de aplicação, bem como exemplos de como realizar configurações adaptáveis para que se possa utilizar esta metodologia nos mais diversos cenários quanto desejado.

Esta metodologia é interessante no que tange o planejamento de eventos, uma vez que ao ser utilizada em diversas aplicações durante ao planejamento de ações pode auxiliar na otimização da utilização de recursos disponíveis.

Dentre as limitações encontradas nesta metodologia, pode-se citar a necessidade de dados de mobilidade, onde a sua ausência impossibilita a aplicação da metodologia. Além disso, para se obter boas recomendações para a organização de eventos voltadas para comunidades específicas, é necessária a criação de comunidades de forma que a comunidade desejada possua diferenças significativas das demais. Caso as comunidades não sejam definidas de forma a que estas possuam diferenças significativas entre si, apesar de não inviabilizar a utilização da metodologia, irá dificultar a proposição de alternativas para o planejamento de eventos de forma a afetar uma comunidade específica.

4.1 Workflow

Nesta seção é apresentada a sequência de etapas da metodologia proposta com a finalidade de obter a influência de eventos em usuários e comunidades.

O desenvolvimento da metodologia proposta foi baseado nas etapas do fluxo de Descoberta de Conhecimento em Base de Dados (KDD), apresentado em detalhes por [40]. Apesar do fluxo desta metodologia se basear no fluxo de [40] no que se refere às etapas, os passos contidos em cada uma das etapas foram adaptados para a obtenção da influência de eventos. Na Figura 4.1 é mostrado o fluxo proposto dividido em etapas e, em cada uma, a lista de passos realizados.

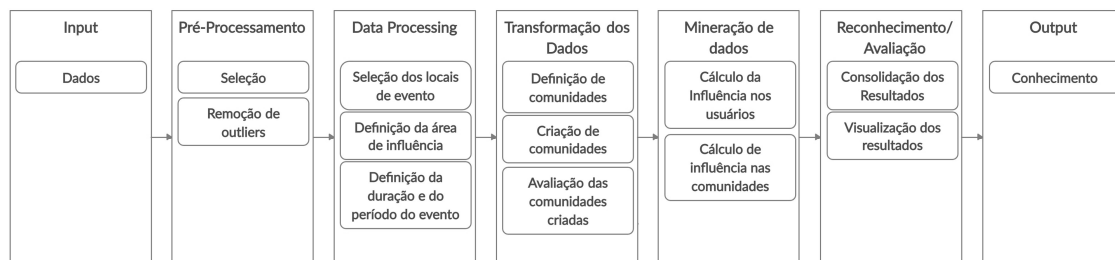


Figura 4.1: Fluxo de atividade para o cálculo da influência de eventos sobre usuários e comunidades.

Os passos apresentados na Figura 4.1 visam um modelo simples, conciso e adaptável, de forma a ser possível sua aplicação em diversos cenários onde se deseja analisar o relacionamento entre usuários, comunidades e eventos. Tais etapas são apresentadas em detalhes a seguir.

4.1.1 *Input*

Os dados de mobilidade são utilizados como *input* para o *workflow* proposto na Figura 4.1, ou seja, são esses dados que irão alimentar o fluxograma para o processamento, o cálculo e a análise da influência de eventos.

Este *workflow* necessita que os registros contendam, no mínimo, as seguintes informações: localização (latitude e longitude; data e hora em que ocorreu a coleta de cada registro; e identificação do usuário do qual o registro foi coletado. Outras informações ainda podem estar presentes no *input* e serem utilizadas para auxiliar na criação de comunidades com outras semânticas, bem como produzir análises aprimoradas do efeito de eventos nestas comunidades, como dados socioeconômicos, sociodemográficos etc.

4.1.2 Pré-Processamento

O pré-processamento é uma etapa de tratamento de dados cuja finalidade é o tratamento do formato dos dados recebidos como *input*, remoção de dados incorretos, mal coletados, fora de padrão ou desinteressantes para o estudo, sendo um passo extremamente importante em virtude da remoção de dados que podem impactar negativamente nos resultados.

Existem diversas técnicas para o tratamento de dados e algumas das mais comuns são: remoção de dados vazios; remoção ou tratamento de *outliers*; normalização; Seleção de dados; *upsampling*; *downsampling*; e geração de dados sintéticos utilizando SMOTE (Synthetic Minority Over-Sampling Technique) ou SMOTE modificado. Essas técnicas podem ser utilizadas para a geração de melhores resultados em casos onde os dados não são suficientes para o experimento ou os dados existentes possuem baixa qualidade, como discutido em [41].

Dentre diversas possíveis fases do pré-processamento, adotou-se a (i) seleção e a (ii) remoção de *outliers*, como detalhadas a seguir.

4.1.2.1 Seleção

A seleção é uma fase do pré-processamento, usada no *workflow*, com o objetivo de realizar a filtragem de dados que não serão utilizados no estudo.

De forma a exemplificar, imagine um cenário onde exista a disponibilidade de dados de mobilidade de todo um país, além de dados socioeconômicos e de prefe-

rência de filmes, porém o estudo possui como público alvo apenas os residentes de um estado com uma análise socioeconômica. Neste caso em tela, os dados deverão ser filtrados e as informações sobre preferência de filmes deverá ser removida.

4.1.2.2 Remoção de *outliers*

A remoção de *outliers* é utilizada para remover dados que destoam dos demais dados daquele tipo, seja devido a um erro de coleta ou por outra razão não identificada.

A título de exemplo, em um *dataset* que possua informações coletadas de radares de velocidade em vias urbanas, devem ser desconsiderados todos os registros coletados com velocidades acima 200 km/h devido a um erro de coleta ou comportamento extremamente não usual de um motorista.

4.1.3 Processamento dos Dados

O processamento dos dados é o passo no qual são definidos os locais dos eventos, o comportamento e as características dos eventos, como data e hora de ocorrência, área de influência, frequência, recorrência, sazonalidade etc.

Na hipótese de eventos que já ocorreram serem o objeto do estudo, se as informações e as características do evento já se encontrarem disponíveis, estas serão utilizadas no processamento de dados. Caso contrário, essas informações e características deverão ser obtidas através de processamento de dados disponíveis acerca do evento para a obtenção dos mesmos.

Já na hipótese de eventos futuros ou planejamento de eventos hipotéticos, as informações também podem ser obtidas a partir de processamento de dados de eventos semelhantes, de processamento dos dados disponíveis utilizados como *input*, pela determinação discricionária do líder do experimento ou definição por especialistas dos eventos analisados.

4.1.3.1 Seleção dos locais de evento

A seleção dos locais de evento é o passo no qual os locais de eventos são escolhidos. Essa escolha pode ser realizada manualmente, aleatoriamente ou baseada em análise de dados.

A seleção manual ocorre quando existe uma lista de escolhas previamente defi-

nida. A seleção de forma aleatória ocorre quando não se tem a priori nenhuma opção de local pré-definida. Já a seleção baseada em análise de dados ocorre quando o local pode ser escolhido utilizando critérios de mobilidade dos dados de usuários disponíveis, respeitando preferencialmente alguma semântica, como estações de trem, estações de metrô, universidades, hospitais, parques, ou outros locais utilizando outras características que atribua significado ao local de evento.

4.1.3.2 Seleção da área de influência de eventos

A seleção da área de influência de eventos é o passo no qual são definidas as áreas em que cada evento irá exercer influência em usuários e comunidades. Esta área é diretamente relacionada com o evento em si, podendo ser maior ou menor que a área de influência de outros eventos.

Na hipótese do evento já ter ocorrido, a área de influência pode ser definida a partir da análise das informações do evento, ou caso não se possua esta informação, pode ser obtida arbitrariamente ou a partir de análise de dados gerais de mobilidade na região em que ocorreu o evento.

Já quanto aos eventos que estão sendo planejados ou eventos hipotéticos, pode-se realizar uma estimativa utilizando como base eventos similares já ocorridos ou definir arbitrariamente variando esta área para uma melhor adaptação de cada evento.

A área de influência pode ser definida em diferentes formatos, como polígonos e círculos com diferentes dimensões. Por simplicidade, sugere-se a utilização de áreas internas a uma circunferência.

4.1.3.3 Seleção do período e duração de eventos

A seleção do período e duração de eventos é a etapa na qual são definidos o período e a duração de cada um dos eventos que estarão no estudo.

De forma similar à definição da área de evento, o período e duração dos eventos que já ocorreram podem ser selecionados a partir da análise do evento em si. Eventos que estão sendo planejados podem ser selecionados arbitrariamente ou através da análise de dados disponíveis de eventos similares.

As características de um evento, como recorrência diária, semanal, alternância de dias e outras características temporais de eventos, também são definidas neste passo.

4.1.4 Transformação de Dados

Neste passo, os dados passados como *input* já tratados são utilizados para a geração de comunidades que se pretende analisar. A transformação de dados pode ser substituída por conhecimento prévio da composição e da configuração das comunidades, obtido em outros estudos.

4.1.4.1 Definição de comunidades

Definição de comunidades é a etapa na qual são estipulados os critérios de criação de comunidades. Esses critérios podem ser baseados em: a) dados geográfico, como região de trabalho, região de moradia etc; b) dados extraídos de uma análise dos dados de mobilidade, como região de trabalho, região de moradia etc; c) dados sociodemográficos, como classe social, idade, gênero, profissão etc; d) outros atributos definidos como relevantes pelos organizadores e especialistas dos eventos que são objetos do estudo.

O critério de criação de comunidades precisa ser definido orientado ao objetivo do estudo, pois os resultados obtidos serão em função das comunidades utilizadas e orientadas a elas.

Após a realização do experimento, existe a possibilidade de se verificar a existência de correlação entre comunidades utilizando as influências obtidas ao longo do tempo. Diversos autores propuseram diferentes formas de agrupar usuários em comunidades - como exemplo, [42] agrupa usuários utilizando uma análise do padrão de mobilidade de cada usuário.

4.1.4.2 Criação de comunidades

A criação de comunidades é o passo no qual as comunidades são efetivamente criadas, ocorrendo a partir das definições dos critérios de criação de comunidades.

Para realizar a criação de comunidades pode ser utilizada a divisão dos usuários por faixas de valores de atributos chave (idade, gênero, classe social etc) ou utilizar algoritmos de classificação para realizar este passo.

Existe uma diversidade de algoritmos de classificação disponíveis na literatura. Entre os mais utilizados, pode-se citar: K-means, DBSCAN, KNN, SVM, Decision Tree e Random Forest. [7] apresenta um *survey* dos algoritmos de classificação mais utilizados e os dividem em categorias de forma a evidenciar os pontos fortes de cada

um deles.

4.1.4.3 Avaliação das comunidades criadas

Avaliação das comunidades criadas é a fase em que é analisado se as comunidades criadas correspondem ao esperado e se existe algum tipo de distorção ocasionada durante a execução dos algoritmos de classificação.

4.1.5 Mineração de dados e cálculo de métricas

Mineração de dados é o passo no qual o cálculo de influência de cada evento sob os usuários e as comunidades são realizados. Todas as definições realizadas são utilizadas neste momento para a realização destes cálculos.

Para realizar o cálculo das métricas de influência, é necessário primeiramente analisar e obter informações dos dados de forma que estes cálculos sejam possíveis. Dentre as informações necessárias, estão a obtenção da Matriz de Influência e o Fator de Influência de cada eventos.

O cálculo da Influência média e Influência relativa, tanto para usuários quanto para comunidades, são realizadas neste passo e são detalhadas a seguir.

4.1.5.1 Cálculo de influência média sobre usuários

O cálculo da influência média sobre os usuários tem por objetivo a obtenção da influência dos eventos ao longo do tempo de observação.

O cálculo da influência média de um evento e_l sobre um usuário u_o ($I_{e_l u_o}$) é realizado utilizando como variáveis a Matriz de Influência (IM), o Fator de Influência do evento e_l (IF_{e_l}) e o período de observação de u_o (Δt_{u_o}). A formulação do cálculo da Influência é apresentada na Equação 4.1.

$$I_{e_l u_o}(\Delta t_{u_o}) = \frac{\sum_{t=1}^{\Delta t_{u_o}} IM_{lot} \cdot IF_{e_l}(t)}{\Delta t_{u_o}} \quad (4.1)$$

4.1.5.2 Cálculo de influência relativa sobre usuários

Quanto ao cálculo da influência relativa sobre os usuários, este também tem por escopo a obtenção da influência dos eventos ao longo do tempo de observação. Todavia, diferentemente do cálculo da influência média sobre os usuários, neste desconsidera-se diferenças no que se refere ao tempo de observação e o ao tempo de

monitoramento dos usuários. O cálculo da $Irel_{e_l u_o}$ é dado pela Equação 4.2.

$$Irel_{e_l u_o}(\Delta t_{u_o}) = \frac{\sum_{t=1}^{\Delta t_{u_o}} IM_{lot} \cdot IF_{e_l}(t)}{\sum_{t=1}^{\Delta t_{u_o}} IF_{e_l}(t)} \quad (4.2)$$

4.1.5.3 Cálculo de influência média sobre comunidades

Cálculo de influência média sobre comunidades visa a obtenção da influência dos eventos nas comunidades ao longo do tempo de observação.

O cálculo da influência de um evento e_l sobre uma comunidade c_m ($Ic_{e_l c_m}$) é realizado utilizando a influência de cada usuário u_q pertencente a comunidade ($I_{e_l u_q}$) e o peso daquele usuário na comunidade c_m (w_{qm}).

Os pesos são atribuídos aos usuários no cálculo da influência de um evento sobre uma comunidade para atender cenários onde existem usuários que representam melhor uma comunidade ou que exerçam uma maior influência sobre os demais indivíduos dessa comunidade.

A formulação do cálculo da influência média sobre comunidades é apresentada na Equação 4.3.

$$Ic_{e_l c_m} = \frac{\sum_{q=1}^Q I_{e_l u_q} \cdot w_{qm}}{\sum_{q=1}^Q w_{qm}} \quad (4.3)$$

4.1.5.4 Cálculo de influência relativa sobre comunidades

Outrossim, o cálculo de influência relativa sobre comunidades tem por escopo a obtenção da influência dos eventos ao longo do tempo de observação, desconsiderando diferenças no tempo de observação e no tempo de monitoramento das comunidades.

O cálculo da influência relativa de um evento e_l sobre uma comunidade c_m ($Icrel_{e_l c_m}$) é realizado utilizando a influência de cada usuário u_q pertencente à comunidade ($Irel_{e_l u_q}$) e o peso daquele usuário na comunidade c_m (w_{qm}).

Assim como no cálculo da medida anterior, os pesos são atribuídos aos usuários no cálculo da influência de um evento sobre uma comunidade para atender cenários onde existem usuários que representam melhor uma comunidade ou que exerçam uma maior influência sobre os demais indivíduos dessa comunidade.

A formulação do cálculo da influência relativa sobre comunidades é apresentada na Equação 4.4.

$$I_{\text{crel}_{e_1 c_m}} = \frac{\sum_{q=1}^Q I_{\text{rel}_{e_1 u_q}} \cdot w_{qm}}{\sum_{q=1}^Q w_{qm}} \quad (4.4)$$

4.1.6 Reconhecimento e Avaliação dos Resultados

Após a conclusão das etapas de coleta e de tratamento dos dados realizados nos passos anteriores, ocorrem o reconhecimento e a avaliação dos resultados, isto é, a informação gerada acerca da influência de cada evento sobre usuários e comunidades são consolidados e apresentados de modo a se obter fácil interpretação pelos dirigentes do estudo.

Destaca-se que a etapa de reconhecimento e de avaliação dos resultados é de suma importância para a validação do estudo e a verificação da validade dos resultados, de forma a identificar supostos erros metodológicos e se os resultados obtidos estão em conformidade com o esperado.

4.1.6.1 Consolidação dos Resultados

Na consolidação dos resultados, os dados gerados para cada usuário e cada comunidade são agrupados individualmente, possibilitando a criação de figuras, gráficos ou qualquer outra forma de representação pretendida dos resultados. Esses resultados individuais podem ser agrupados e agregados de forma a ilustrar um único resultado global para todos usuários e todas comunidades.

4.1.6.2 Visualização dos Resultados

Na etapa de visualização dos resultados realiza-se o processo de geração de gráficos e estatísticas a partir dos dados agregados na fase de consolidação dos resultados. Ressalta-se que as figuras, as estatísticas e os gráficos gerados são fundamentais para a extração de conhecimento e a aplicação em casos reais.

4.1.7 Output

Após todo o processamento e análise dos dados, é gerado como *output* o conhecimento que poderá ser utilizado em aplicações em casos reais, possibilitando assim a otimização da utilização de recursos ou minimização de transtornos, quando

possível, na execução desses eventos.

4.2 Exemplo Aplicado

O objetivo da metodologia proposta é ser genérica o bastante para ser aplicável em diversos cenários onde se deseja analisar a relação entre usuários, eventos e comunidades.

Dentre os cenários possíveis de aplicação desta metodologia, pode-se mencionar: a) o planejamento de eventos de rua, de forma a minimizar o impacto no tráfego diário; b) o planejamento de exposição de peças de marketing de modo a maximizar o alcance de grupos de público alvo ou minimizar os custos a partir da seleção de pontos e horários estratégicos de exibição; c) a organização e planejamento de eventos fora de pontos de grande movimentação porém de fácil acesso a usuários.

Para exemplificar a aplicação desta metodologia de uma forma simples e demonstrar os valores que são assumidos de influência para usuários e comunidades, é apresentado a seguir um exemplo com valores hipotéticos.

Sejam os usuários u_1 , u_2 e u_3 os usuários que se deseja monitorar e sejam as comunidades $c_1 = \{u_1, u_2\}$, $c_2 = \{u_1, u_3\}$ e $c_3 = \{u_2, u_3\}$. Cada usuário u_o foi monitorado ao longo de 24 horas de forma ininterrupta e cada coleta de posicionamento foi realizada em intervalos de 10 minutos cada, obtendo-se ao final do monitoramento 144 registros para cada usuário, ou 432 registros ao todo.

Suponha dois eventos e_1 e e_2 , sendo ambos sobre a lentidão de tráfego na cidade de São Paulo (SP - Brasil) e com a mesma área de influência. Os cenários de e_1 e e_2 são descritos a seguir:

1. Em e_1 , o Fator de Influência (IF_{e_1}) é dado pela função $IF_{e_1}(t) = 1$, apresentado na Figura 4.2, e ocorre durante dois períodos: o primeiro com início às 08:00 e fim às 10:00; e o segundo com início às 17:00 e fim às 20:00;
2. Em e_2 , o Fator de Influência (IF_{e_2}) é dado pela função apresentada na Figura 4.3 e o evento ocorre em um período único com início às 07:00 e fim às 20:00.

As funções que representam o Fator de Influência (IF) para ambos eventos foram geradas a partir de uma simplificação da função apresentado em [43].

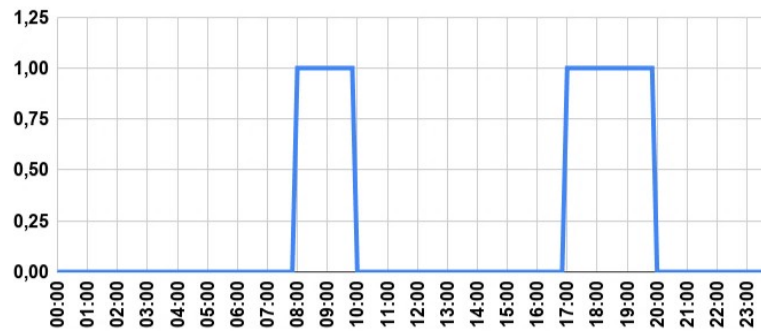


Figura 4.2: Representação do Fator de Influência para o evento e_1 (IF_{e_1}) (Eixo Y) em função do tempo (Eixo X).

Após o monitoramento dos usuários u_1 , u_2 e u_3 , foram feitas observações dos instantes de tempo que cada um deles estiveram dentro da área de influência dos eventos. Essas observações são apresentadas na Tabela 4.1 e, a partir delas, gera-se a Matriz de Influência (IM).

Tabela 4.1: Horário de entrada e saída da área de influência dos eventos e_1 e e_2 pelos usuários u_1 , u_2 e u_3 .

	Manhã		Noite	
	Entrada	Saída	Entrada	Saída
u_1	08:00	09:00	18:00	20:00
u_2	09:00	10:00	17:00	18:00
u_3	06:00	12:00	12:00	22:00

Após a definição dos eventos ter sido realizada e os dados dos usuários terem sido coletados e processados, foi realizado o cálculo das influências médias e relativas dos eventos e_1 e e_2 sobre os usuários u_1 , u_2 e u_3 .

As Figuras 4.4 (a), (b) e (c) exibem a Influência Média ao longo do tempo do evento e_1 para os usuários u_1 , u_2 e u_3 respectivamente. Nestas figuras, as linhas vermelhas representam a Influência Média para cada usuário em cada instante de tempo t , do instante $t = 0$ ao último instante do dia. A linha azul representa o Fator de Influência $IF_{e_1}(t)$.

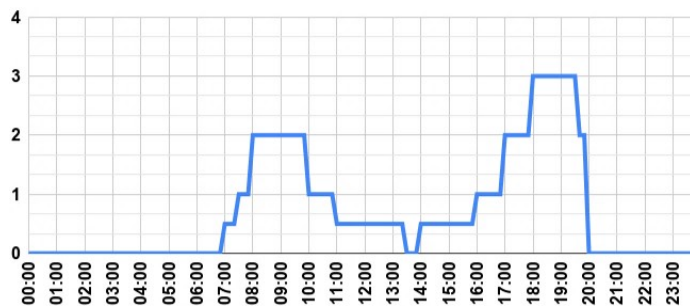
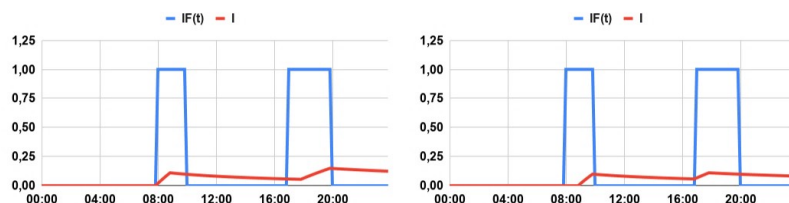
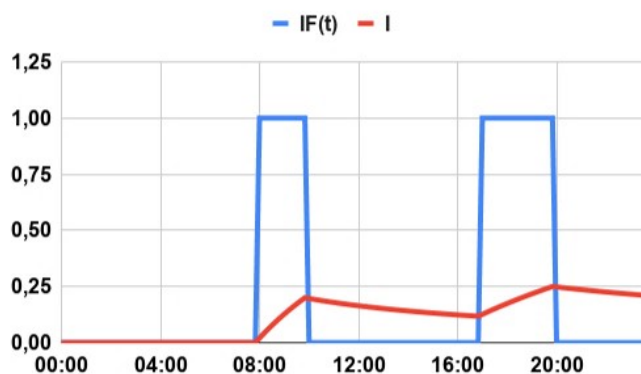


Figura 4.3: Representação do Fator de Influência para o evento e_2 (IF_{e_2}) (Eixo Y) em função do tempo (Eixo X).



(a)

(b)

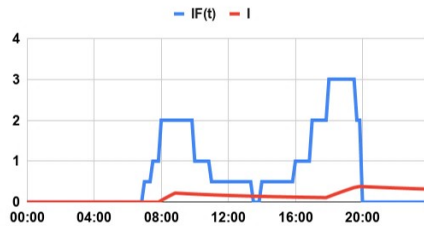


(c)

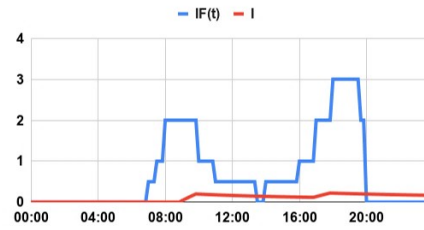
Figura 4.4: Influência Média e Fator de Influência (Eixo Y) do evento e_1 em função do tempo (Eixo X) sobre os usuários u_1 (a), u_2 (b) e u_3 (c).

Ao fim das 24 horas de monitoramento, foram obtidos os seguintes valores de Influência Média do evento e_1 : para o usuário u_1 , obteve-se o valor 0.125; para o usuário u_2 , obteve-se o valor 0.08333; e para o usuário u_3 , obteve-se o valor 0.208333.

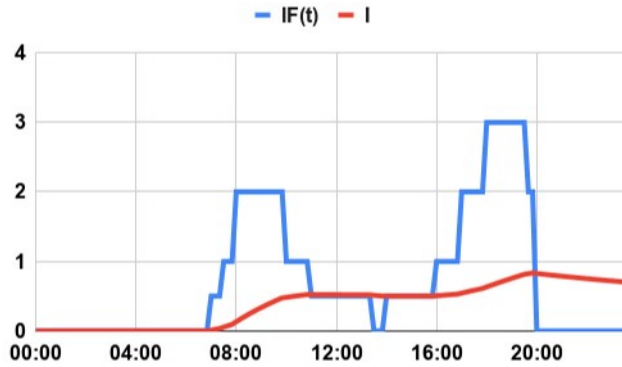
As Figuras 4.5 (a), (b) e (c) exibem a Influência Média ao longo do tempo do evento e_2 para os usuários u_1 , u_2 e u_3 respectivamente. Nestas figuras, as linhas vermelhas representam a Influência Média para cada usuário em cada instante de tempo t , do instante $t = 0$ ao último instante do dia. Já a linha azul representa o Fator de Influência $IF_{e_2}(t)$.



(a)



(b)



(c)

Figura 4.5: Influência e Fator de Influência Média (Eixo Y) do evento e_2 em função do tempo (Eixo X) sobre os usuários u_1 (a), u_2 (b) e u_3 (c).

Ao fim das 24 horas de monitoramento, foram obtidos os seguintes valores de Influência Média do evento e_2 : para o usuário u_1 , obteve-se o valor 0.319444; para o usuário u_2 , obteve-se o valor 0.1666; e para o usuário u_3 , obteve-se o valor 0.69444.

Após o cálculo da Influência Média ter sido realizado para cada usuário, é realizado o cálculo da Influência Relativa (I_{rel}). As Figuras 4.6 (a), (b) e (c) apresentam a Influência Relativa de e_1 para os usuários u_1 , u_2 e u_3 , respectivamente. Nessas figuras, as linhas verdes representam a Influência Relativa para cada usuário em cada instante de tempo t , iniciando em $t = 0$ indo até o último instante do dia. Já as linhas azuis representam o Fator de Influência para cada instante de tempo t . Os valores obtidos para u_1 , u_2 e u_3 foram 0.6, 0.4 e 1.0 respectivamente.

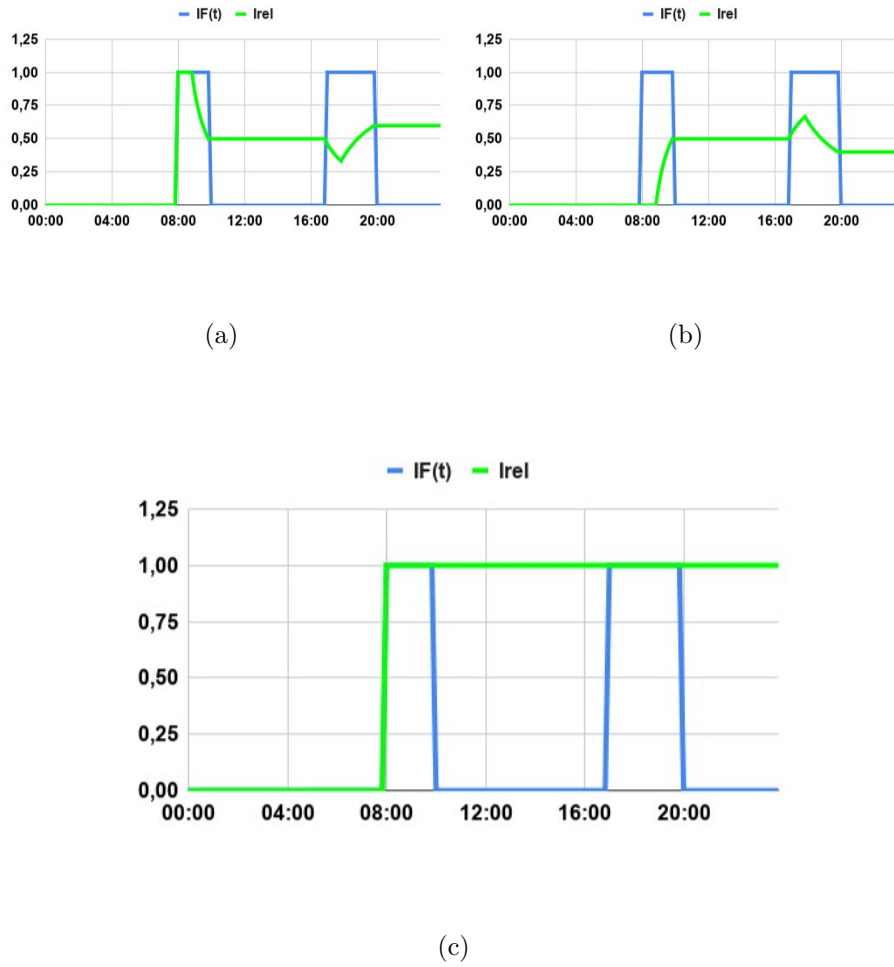


Figura 4.6: Influência Relativa e Fator de Influência (Eixo Y) em função do tempo (Eixo X) para o evento e_1 sobre os usuários u_1 (a), u_2 (b) e u_3 (c).

As Figuras 4.7 (a), (b) e (c) apresentam a Influência Relativa de e_2 para os usuários u_1 , u_2 e u_3 , respectivamente. Nessas figuras, as linhas verdes representam a Influência Relativa para cada usuário em cada instante de tempo t , iniciando em $t = 0$ indo até o último instante do dia. As linhas azuis representam o Fator de Influência para cada instante de tempo t . Os valores obtidos para u_1 , u_2 e u_3 foram

0.46, 0.24 e 1.0 respectivamente.

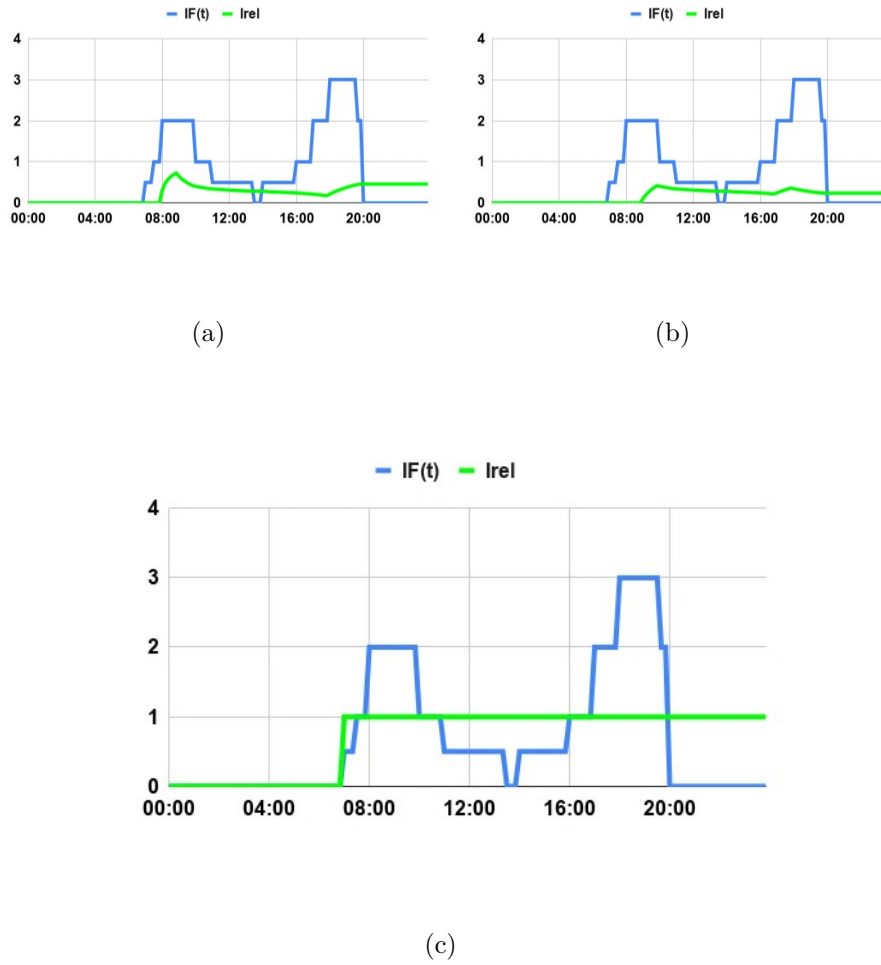


Figura 4.7: Influência Relativa e Fator de Influência (Eixo Y) em função do tempo (Eixo X) para o evento e_2 sobre os usuários u_1 (a), u_2 (b) e u_3 (c).

É importante ressaltar que, como a Influência relativa é determinada apenas pelo período de tempo no qual o evento estava efetivamente ativo, a variação do valor desta medida ocorre apenas nos períodos em que os eventos estavam ativos e, assim, exercendo influência sobre os usuários.

Após os cálculos da Influência Média e Influência Relativa terem sido obtidos para os usuários u_1 , u_2 e u_3 , calcula-se a Influência Média e a Influência Relativa para as comunidades c_1 , c_2 e c_3 .

As Figuras 4.8 (a), (b) e (c) apresentam a Influência de e_1 para as comunidades c_1 , c_2 e c_3 , respectivamente. Nessas figuras, linhas vermelhas representam a Influência Média para cada comunidade em cada instante de tempo t , iniciando em $t = 0$ indo

até o último instante do dia. As linhas azuis representam o Fator de Influência do evento e_1 para cada instante de tempo t . Os valores obtidos para c_1 , c_2 e c_3 foram 0.1042, 0.1666 e 0.1458, respectivamente.

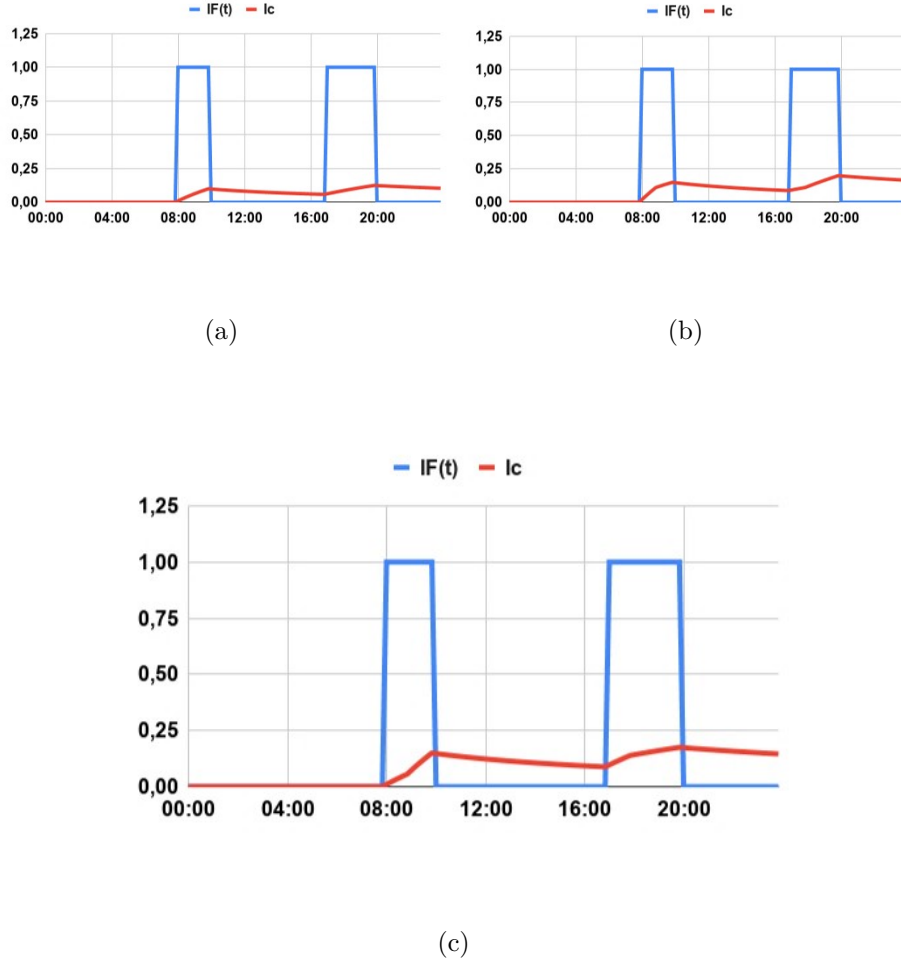


Figura 4.8: Influência Média e Fator de Influência (Eixo Y) em função do tempo (Eixo X) para o evento e_1 sobre as comunidades c_1 (a), c_2 (b) e c_3 (c).

As Figuras 4.9 (a), (b) e (c) apresentam a Influência Média de e_2 para as comunidades c_1 , c_2 e c_3 , respectivamente. Nessas figuras, as linhas vermelhas representam a Influência Média para cada comunidade em cada instante de tempo t , iniciando em $t = 0$ indo até o último instante do dia. As linhas azuis representam o Fator de Influência do evento e_2 para cada instante de tempo t . Os valores obtidos para c_1 , c_2 e c_3 foram 0.2431, 0.5069 e 0.4306, respectivamente.

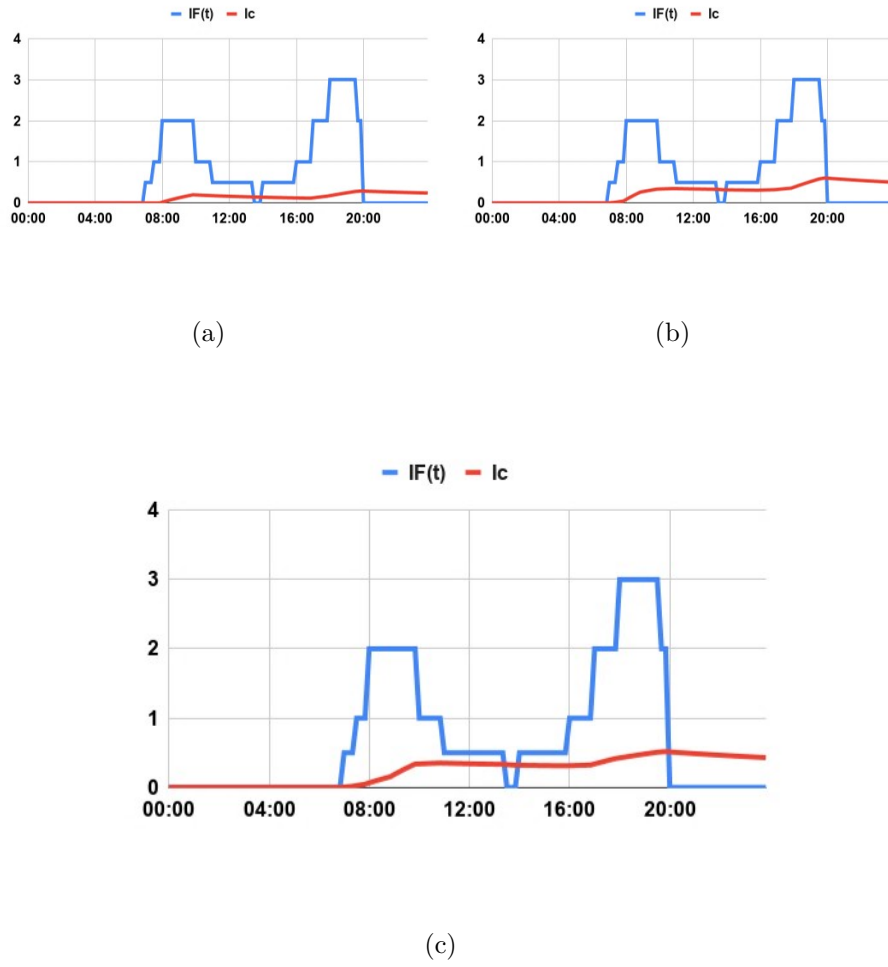


Figura 4.9: Influência e Fator de Influência (Eixo Y) em função do tempo (Eixo X) para o evento e_2 sobre as comunidades c_1 (a), c_2 (b) e c_3 (c).

As Figuras 4.10 (a), (b) e (c) apresentam a Influência Relativa de e_1 para as comunidades c_1 , c_2 e c_3 , respectivamente. Nessas figuras, as linhas verdes representam a Influência para cada comunidade em cada instante de tempo t , iniciando em $t = 0$ indo até o último instante do dia. As linhas azuis representam o Fator de Influência do evento e_1 para cada instante de tempo t . Os valores obtidos para c_1 , c_2 e c_3 foram 0.5, 0.8 e 0.7, respectivamente.

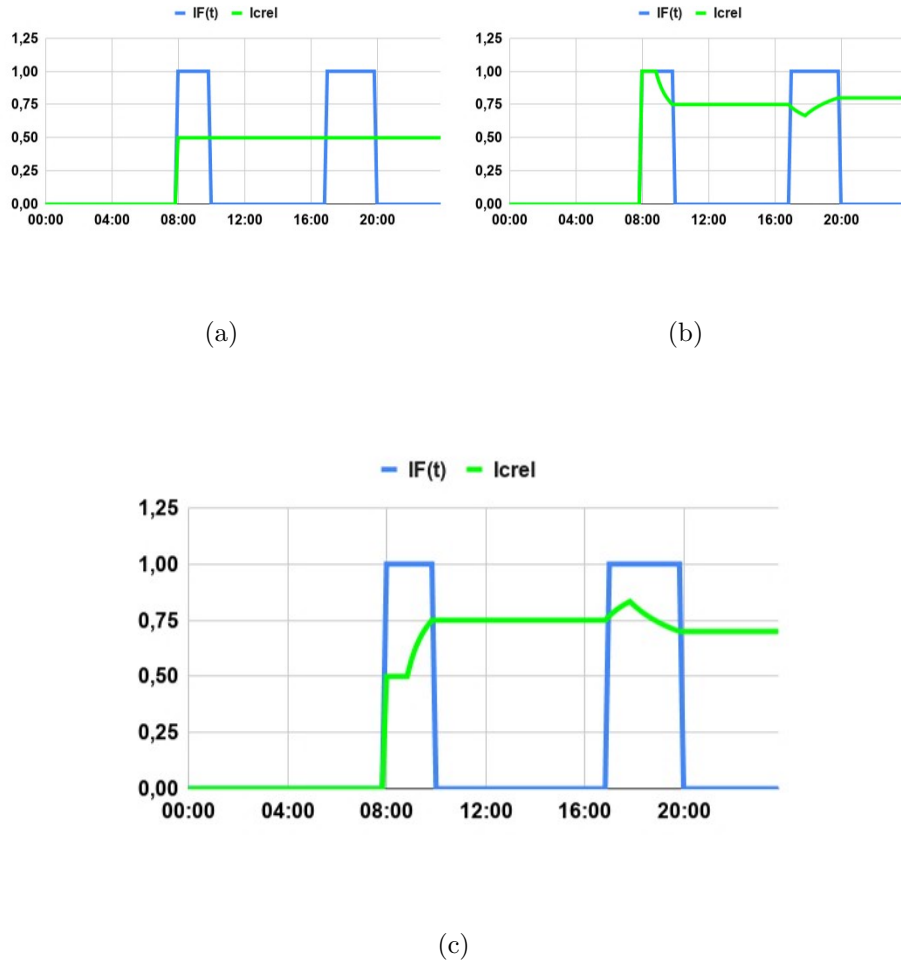
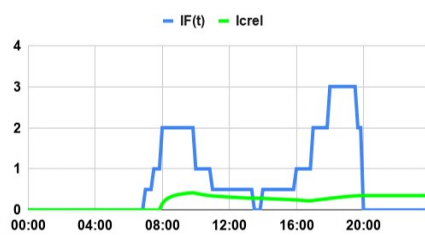
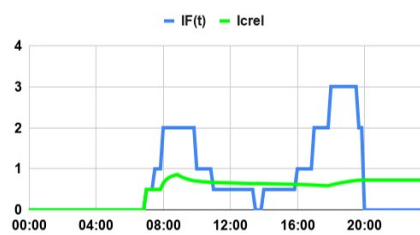


Figura 4.10: Influência Relativa e Fator de Influência (Eixo Y) em função do tempo (Eixo X) para o evento e_1 sobre as comunidades c_1 (a), c_2 (b) e c_3 (c).

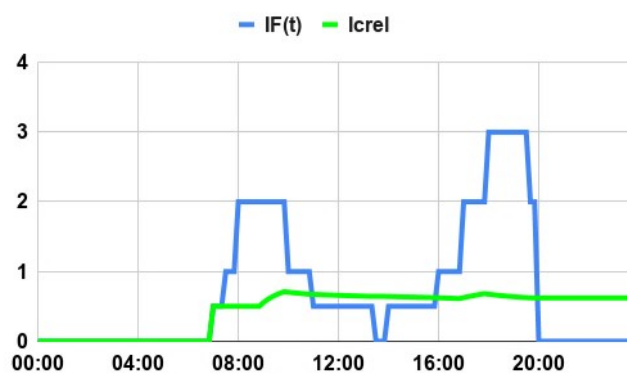
As Figuras 4.11 (a), (b) e (c) apresentam a Influência Relativa de e_2 para as comunidades c_1 , c_2 e c_3 , respectivamente. Nessas figuras, as linhas verdes representam a Influência para cada comunidade em cada instante de tempo t , iniciando em $t = 0$ indo até o último instante do dia. As linhas azuis representam o Fator de Influência do evento e_2 para cada instante de tempo t . Os valores obtidos para c_1 , c_2 e c_3 foram 0.35, 0.73 e 0.62, respectivamente.



(a)



(b)



(c)

Figura 4.11: Influência Relativa e Fator de Influência (Eixo Y) em função do tempo (Eixo X) para o evento e_2 sobre as comunidades c_1 (a), c_2 (b) e c_3 (c).

5. Experimentação

Neste capítulo são apresentadas e detalhadas as fontes de dados utilizadas, bem como os passos realizados durante a experimentação, ou seja, ao longo do processo do experimento.

Durante a execução do processo de experimentação utilizou-se a linguagem de programação *Python*, através da aplicação *Jupyter Notebook* integrada com o gerenciador de pacotes *Anaconda*. Essas ferramentas foram utilizadas em conjunto com diversas bibliotecas de *data science* disponíveis para a linguagem *Python* para a realização dos cálculos e geração da visualização de dados e visualização de resultados.

5.1 *Datasets*

Nesta seção são apresentadas e detalhadas as fontes de dados utilizadas na experimentação. Foram utilizados dois *datasets* bastante difundidos na literatura para demonstrar a aplicabilidade da metodologia proposta.

5.1.1 Geolife

O *dataset* Geolife [3] é formado por dados reais de mobilidade coletados de 182 usuários durante o período entre 2007 e 2012, contendo um total de 24.876.986 registros. Cada um desses registros possuem dados sobre: data/hora UTC do registro; Latitude; Longitude; Altitude; Modal de transporte utilizado; e ID do usuário. Apesar de o *dataset* possuir informação sobre o modal utilizado, apenas uma fração dos registros possui esta informação (21,8% - 5.427.117 registros).

A maior parte dos registros do Geolife está concentrada na região da grande Pequim. Essa característica é demonstrada na Figura 5.1, através de um mapa de

calor utilizando todos os registros do *dataset*, e na Figura 5.2, que ilustra um mapa de calor após um zoom da região da grande Pequim.

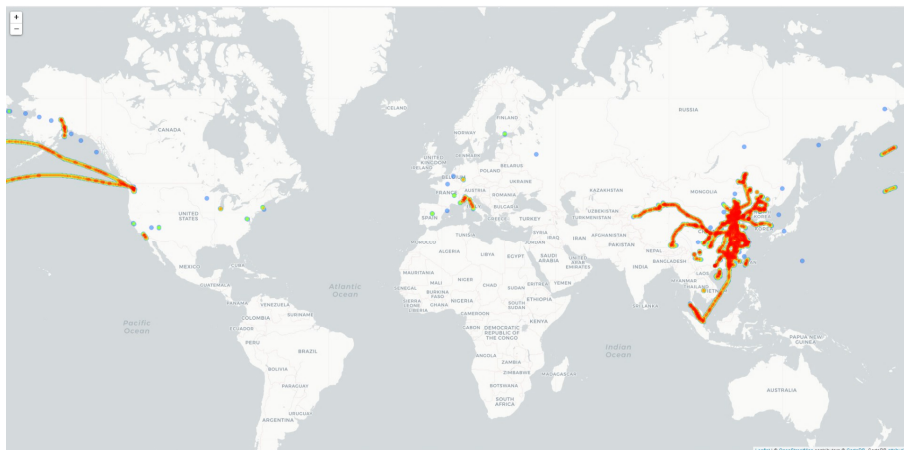


Figura 5.1: Mapa de calor dos registros presentes no *dataset* sobre mapa mundi.

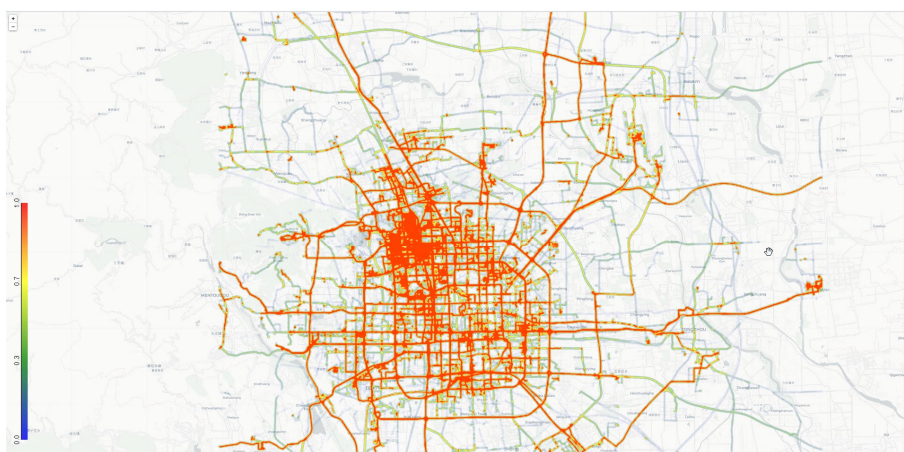


Figura 5.2: Mapa de calor dos registros sobre o mapa da região da grande Pequim.

Outrossim, a distribuição temporal dos registros no Geolife não ocorre de forma

uniforme ao longo do tempo, como bem ilustrado na Figura 5.3. Por outro lado, a distribuição de registros por dia da semana ocorre de forma mais uniforme, como visto na Figura 5.4. De forma similar, a distribuição de usuários que tiveram coleta de dados por dia da semana também se comportou de forma uniforme, como visto na Figura 5.5.

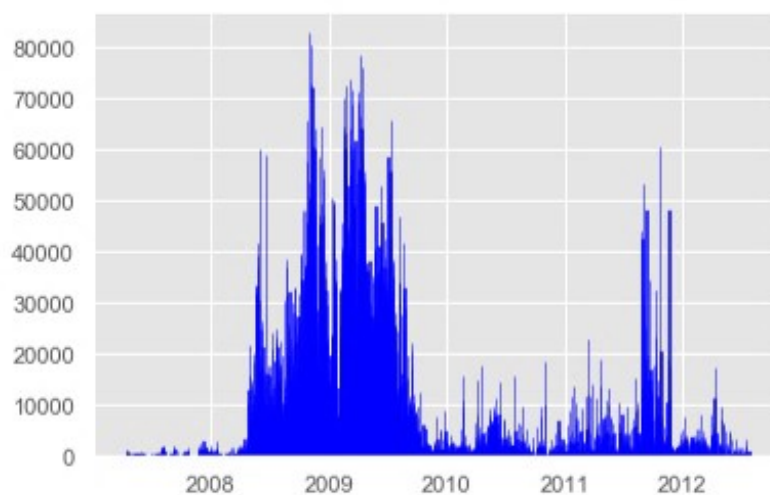


Figura 5.3: Distribuição de registros ao longo do tempo.

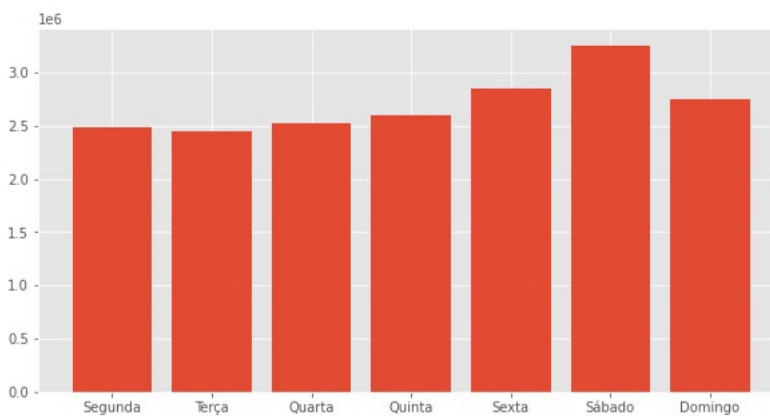


Figura 5.4: Distribuição de registros por dia da semana.

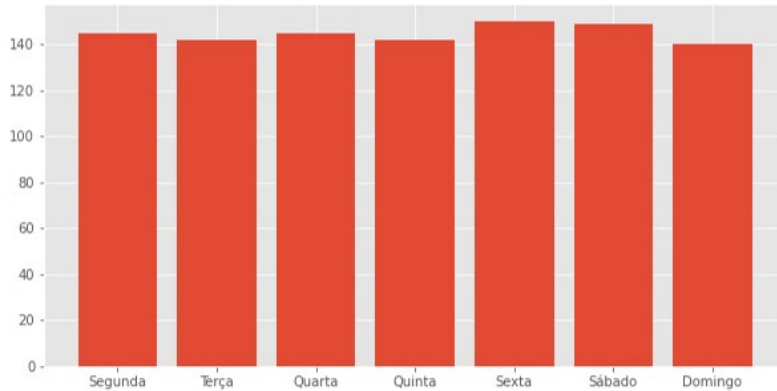


Figura 5.5: Distribuição de usuários ativos por dia da semana.

5.1.2 Cabspotting

O *dataset* Cabspotting [4] é formado por dados reais de mobilidade coletados de 536 táxis de São Francisco (EUA), entre 17/05/2008 e 10/06/2008, contendo um total de 11.219.419 registros. Cada um desses registros possui dados sobre: data/hora do registro; Latitude; Longitude; Estado (livre ou com passageiro); e ID do usuário.

Em relação à distribuição espacial dos registros dos táxis, os dados estão espalhados na região metropolitana de São Francisco, o que inclui, além da cidade de São Francisco, as cidades de Berkely, Oakland e Freemont. Portanto, os dados estão espalhados na região metropolitana de São Francisco, conforme apresentado no mapa de calor exibido na Figura 5.6.

Já em relação à distribuição temporal dos registros no *dataset*, os dados estão distribuídos conforme as seguintes figuras: (a) Figura 5.7, na qual é apresentada a distribuição dos registros por dia da semana; (b) Figura 5.8, na qual é apresentada a distribuição dos registros por data e; (c) Figura 5.9, na qual é apresentada a distribuição dos registros por horário.

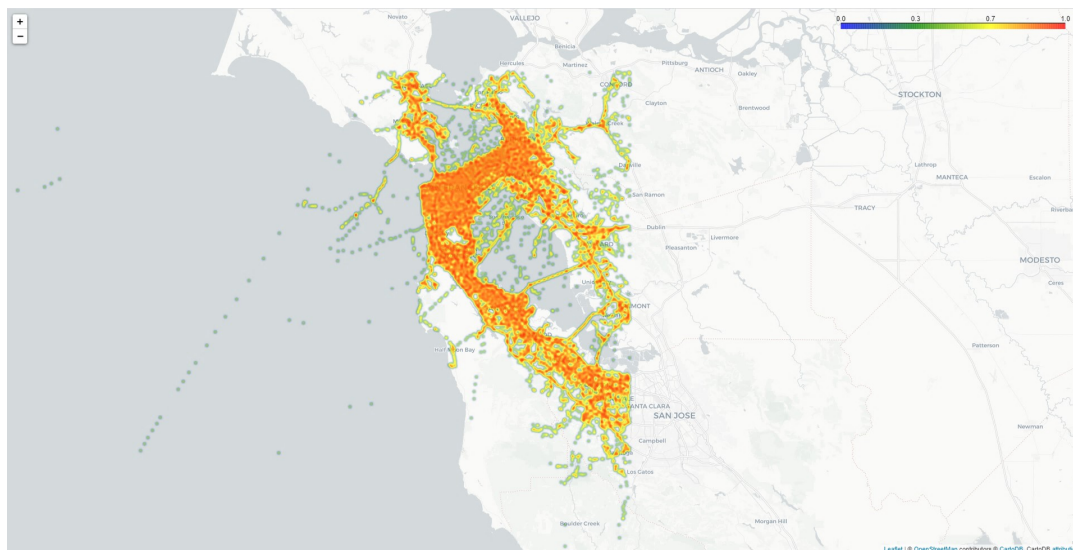


Figura 5.6: Distribuição espacial dos registros utilizando mapa de calor sobre o mapa da cidade de São Francisco (EUA).

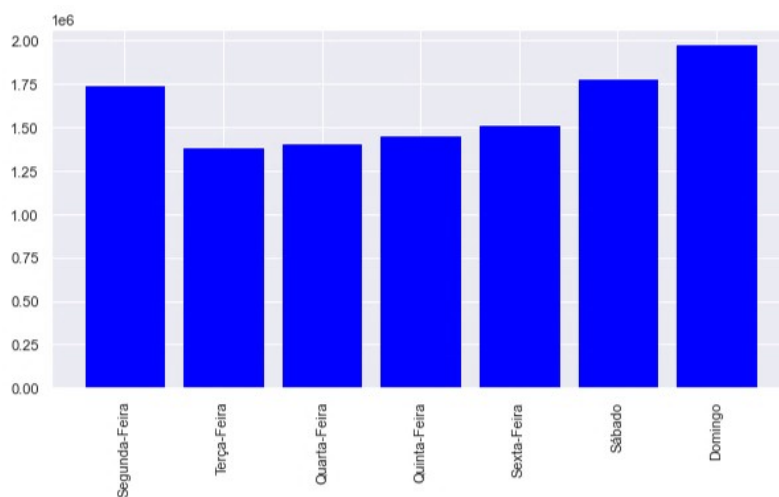


Figura 5.7: Distribuição temporal dos registros por dia da semana.

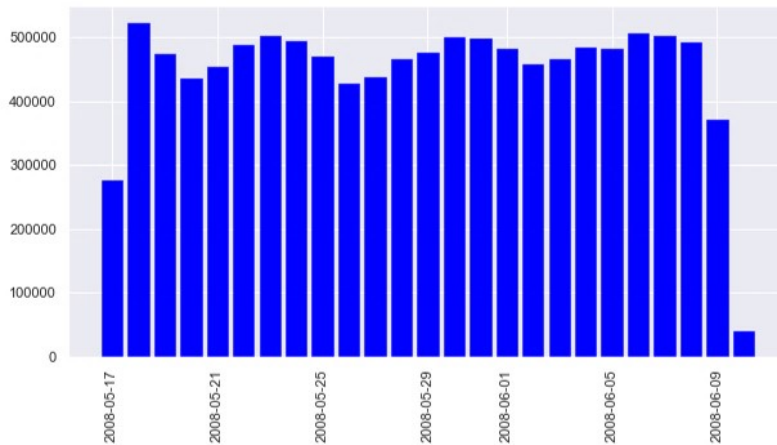


Figura 5.8: Distribuição temporal dos registros por data.

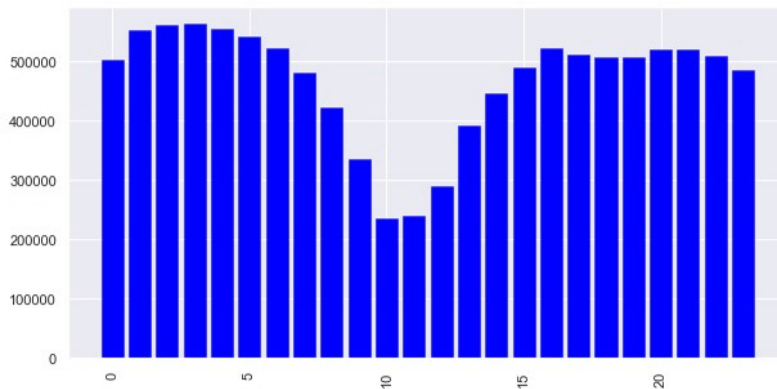


Figura 5.9: Distribuição temporal dos registros por horário.

5.2 Experimento 1

Neste primeiro experimento são apresentados os desenvolvimentos dos passos e etapas necessários para a aplicação de um caso particular da metodologia proposta. O desenvolvimento deste caso particular foi apresentado e publicado em maiores detalhes em [44]. Foi utilizado o *dataset* GEOLIFE [3] e suas características já

foram apresentadas na Seção 5.1.1.

Como primeiro passo do experimento, foi realizada uma filtragem dos dados para que estes estivessem compreendidos na região de interesse, compreendida na região da grande Pequim, realizando a delimitação espacial dos dados entre as coordenadas 39,800000 e 40,200000 de latitude e entre as coordenadas 116,100000 e 116,800000 de longitude. Os dados resultantes dessa filtragem foram apresentados na Figura 5.2, na qual as coordenadas médias resultantes foram de 39,97382 e de 116,3613 para latitude e longitude, respectivamente. Após esta filtragem, os dados resultantes foram no total de 18.750.290 registros.

A distribuição temporal dos dados é apresentada na Figura 5.3, na qual é possível notar que essa distribuição não ocorre de forma homogênea ao longo do tempo. Por esse motivo, este experimento foi realizado considerando que os eventos são recorrentes conforme o dia da semana. A distribuição dos registros por dia da semana é apresentada na Figura 5.4. Além dessa distribuição, é importante também ressaltar a distribuição dos usuários ao longo da semana. Esta distribuição é apresentada na Figura 5.5.

5.2.1 Classificação dos usuários em comunidades

A primeira etapa do experimento consistiu em gerar comunidades. Como dito na Seção 3.2.4, comunidades podem ser definidas de diversas formas e com diversos significados. Para a criação das comunidades, neste experimento, utilizou-se como critério de classificação de usuários o atributo “região de moradia” dos usuários presentes no *dataset*. Devido ao fato dessa informação não estar contida de maneira explícita no *dataset*, os dados disponíveis foram tratados e foi utilizado um algoritmo de agrupamento para a obtenção desta informação do *dataset* utilizando a heurística apresentada a seguir.

- *Um usuário X pertence a uma comunidade Y se este está em uma determinada localidade Z em determinado período de tempo T , representando o horário de recolhimento do usuário em sua residência.*

Foi definido o período T sendo o intervalo entre às 00:00h e 00:30h no horário local, sendo utilizados todos os dados disponíveis para esta faixa de tempo para a realização da classificação dos usuários em comunidades.

Como em toda heurística, os resultados obtidos não são precisos, sendo apenas aproximações da realidade. Neste cenário específico, existe a possibilidade de que usuários que trabalham em horário noturno sejam classificados em comunidades diferentes do correto. Além disso, existe a possibilidade de que usuários troquem de residência ao longo do tempo e desta forma sejam classificados em mais de uma comunidade, assim como desenvolvido em [2, 26], que descrevem a possibilidade de pertencimento a mais de uma comunidade. Como a metodologia proposta visa permitir este tipo de relacionamento, onde usuários podem pertencer a múltiplas comunidades, não é realizado nenhum tratamento para esta questão neste experimento. Outros experimentos, ao contrário deste, pode realizar a classificação de usuários em comunidades onde cada usuário pertença a apenas uma comunidade.

caso seja o objetivo do caso específico estudado,

Para realizar a classificação dos usuários em comunidades, foi utilizado o algoritmo de clusterização K-Means. Os dados de localização, latitude e longitude, filtrados conforme explicado anteriormente, foram utilizados como *input* para o K-Means. A escolha do K-Means foi devido à excelente performance computacional e à grande quantidade de dados a ser processada. Outros algoritmos de clusterização, como o DBSCAN, podem levar um tempo maior, em até 12 vezes quando comparado com K-Means para a obtenção de *clusters* com precisão similar [16].

Como explicado na Seção 2.1.2.3, o K-Means necessita da definição do número K de *clusters* a priori de sua execução. Neste experimento optou-se pelo método Silhouette, devido a sua ampla adoção, boa precisão e facilidade de uso.

De forma a exemplificar possíveis configurações espaciais de comunidades geradas através do K-Means para os dados disponíveis neste *dataset*, é apresentada a Figura 5.10 ilustrando as três configurações de *clusters*.

Para o cálculo do Silhouette Score, foi variado o parâmetro K de 2 a 15 e os resultados para cada um dos K s são apresentados na Figura 5.11(a). É importante notar que, quanto melhor o agrupamento, maior o valor obtido no cálculo do Silhouette score. Desta forma, será utilizado o valor $K = 7$ devido a este ser o melhor resultado obtido. A distribuição espacial das comunidades formadas utilizando o K-Means para $K = 7$ são exibidas na Figura 5.11(b).

Devido ao critério de classificação utilizado para o agrupamento dos usuários com o K-Means e por alguns usuários por não possuírem registros no período selecionado, estes não foram classificados em nenhuma das sete comunidades criadas pelo

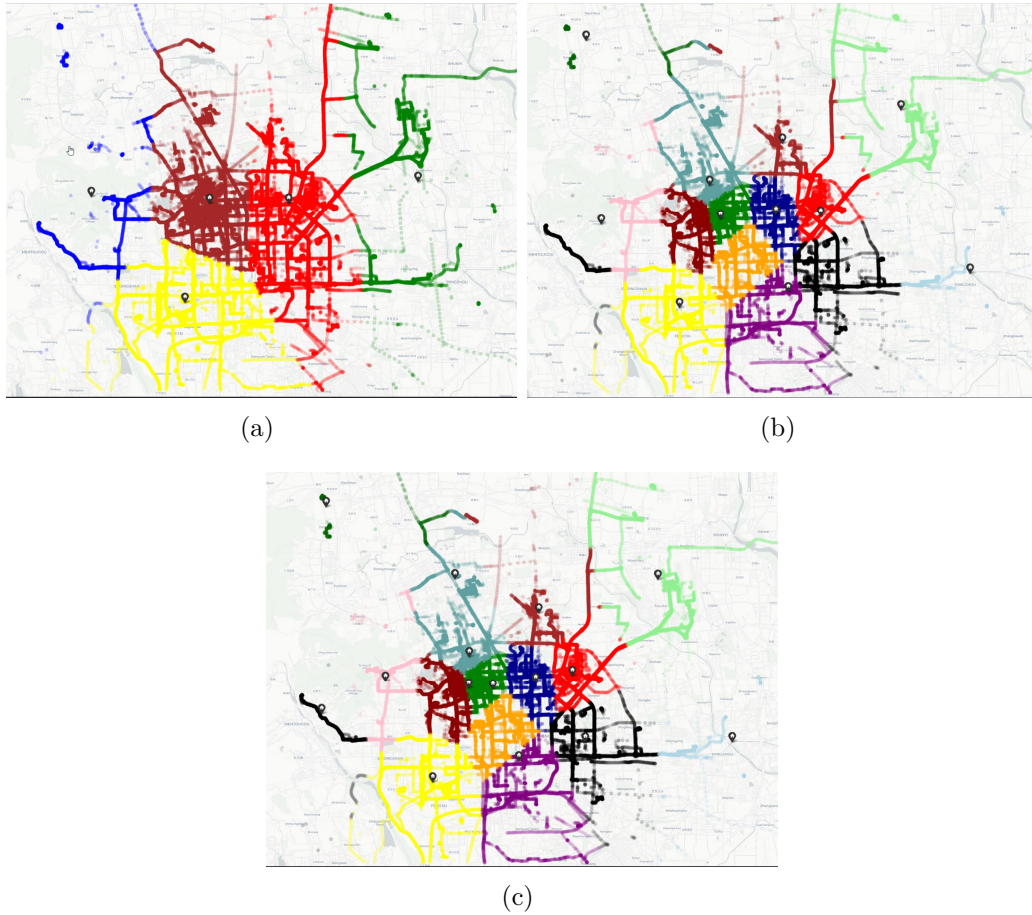


Figura 5.10: *Clusters* dos locais de residência das comunidades para K igual a 5, 11 e 15.

K-Means. De forma a utilizar os dados destes usuários e realizar comparações destes com os demais usuários, foi criada uma oitava comunidade com todos os usuários que não se enquadraram nas comunidades obtidas utilizando o K-Means. O quantitativo do número de usuários por comunidade é apresentada na Tabela 5.1, onde C0 é a comunidade dos usuários sem registros noturnos.

C1	C2	C3	C4	C5	C6	C7	C0
108	51	23	41	36	59	32	55

Tabela 5.1: Número de usuários por comunidade.

5.2.2 Definição de Eventos

Após o tratamento dos dados de usuários e a criação de comunidades utilizando o critério de região de moradia obtido a partir da análise de dados de mobilidade de usuários disponíveis, são realizadas a escolha e as definições dos eventos que serão analisados no experimento.

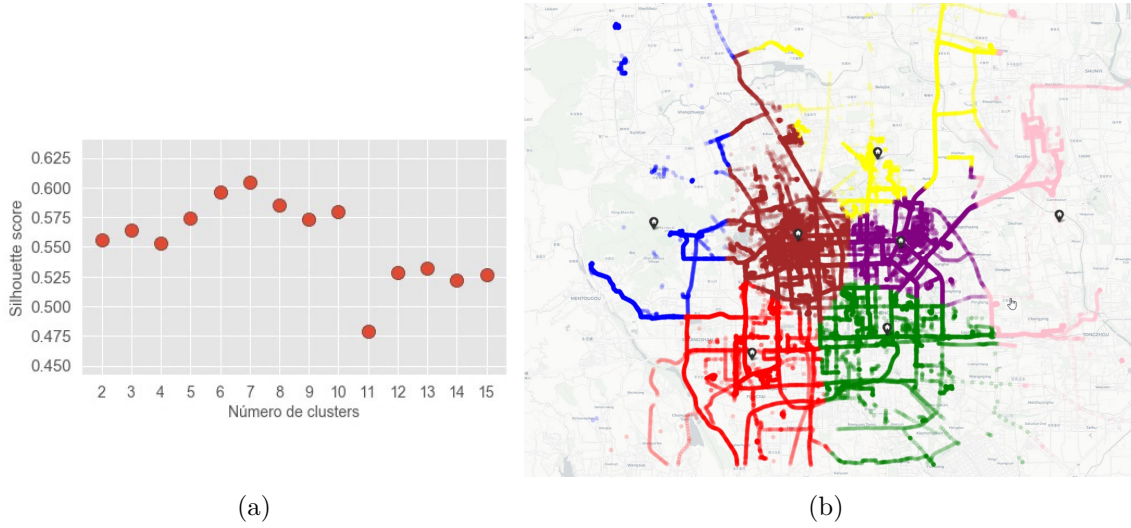


Figura 5.11: *Silhouette score* em função de K (a) e Distribuição espacial do local de residência das comunidades para K ótimo ($K = 7$) (b).

As características de cada evento e_i serão definidas conforme os critérios a seguir:

1. A área de cada evento e_i é definida como a área de uma circunferência de raio de 100 metros, onde o centro da circunferência é o epicentro do evento.
2. A distância entre o centro de um evento e a posição de um usuário u_o é obtida utilizando a distância geodésica de u_o a e_i para cada instante t .
3. O Fator de Influência para cada evento neste experimento foi definido como $IF = 1$, devido ao não conhecimento prévio da intensidade, ao longo do tempo, dos eventos selecionados. Além disso, optou-se pela utilização de $IF = 1$ com o objetivo de realizar a comparação de eventos de maneira linear ao longo do tempo, de forma a não adicionar pesos, maiores ou menores, a intervalos específicos do tempo de duração destes eventos.
4. Por fim, o cálculo de influência do evento e_i sobre cada usuário será realizado utilizando um caso particular da Equação 4.1, sendo ela:

$$\text{influencia}_{ij} = \frac{\sum_{k=0}^n u_k(j) \cdot \text{atingido}(i, u_k(j))}{\sum_{k=0}^n u_k(j)}$$

$$\text{atingido}(i, k) = \begin{cases} 0, & \text{if } \text{distancia}(E_i, u_k(j)) > \text{raio}_i \\ 1, & \text{if } \text{distancia}(E_i, u_k(j)) \leq \text{raio}_i \end{cases}$$

Para as localizações dos eventos a serem analisados, foram escolhidos onze locais onde se possui uma expectativa de constante movimentação de pessoas e veículos.

Além disso, buscou-se a escolha de locais com alguma semântica atrelada a eles, como estações de trem, metrô, clubes, universidades etc. Os locais selecionados para a ocorrência destes eventos são:

1. **Via de acesso ao aeroporto internacional de Pequim** (40.06519, 116.5882);
2. **5th Ring Road, próximo ao museu de ciência e tecnologia da China** (40.02228, 116.41758);
3. **Estação de trem “Beijing West”** (39.89716, 116.32107);
4. **Centro de Tênis de Pequim** (39.85218, 116.4138);
5. **Universidade de Tecnologia de Pequim** (39.8726, 116.48573);
6. **Clube Internacional de Golfe de Pequim** (39.959684, 116.234717);
7. **3 Qing Long Qiao Hao Jie** (40.004911, 116.268255);
8. **Clube de Golfe de Pequim Yanxi** (39.946325, 116.215314);
9. **Laoshan Velodrome** (39.914265, 116.211974);
10. **Parque “Mundo”de Pequim** (39.812638, 116.279310);
11. **Estação de trem “Qinghuayuan”** (39.982417, 116.339780).

Estes locais são apresentados espacialmente na Figura 5.12 e cada centro de evento está identificado sobre o mapa de calor dos registros dos usuários sobre Pequim.

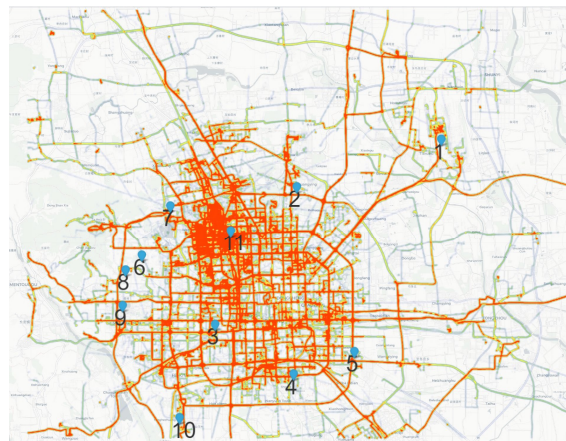


Figura 5.12: Localização geográfica dos locais de ocorrência dos eventos mostrados no *Heat Map* dos dados no mapa de Pequim.

5.2.3 Apresentação dos Resultados

Após a realização de todos os passos de análise e tratamento de dados, definição e criação de comunidades e eventos, é realizado o cálculo da influência dos eventos sobre as comunidade e busca-se apresentar os resultados de forma clara e objetiva. Por simplicidade, daqui em diante o termo “influência” será usado para designar a medida de “influência média”, apresentada na Seção 4.1.5.3.

Os resultados obtidos, também apresentados em [44], são apresentados através das Figuras 5.13, 5.14, 5.15, 5.16, 5.17, 5.18 e 5.19, representando os resultados obtidos nas segundas, terças, quartas, quintas, sextas, sábados e domingos, respectivamente, de forma a facilitar a compreensão dos resultados.

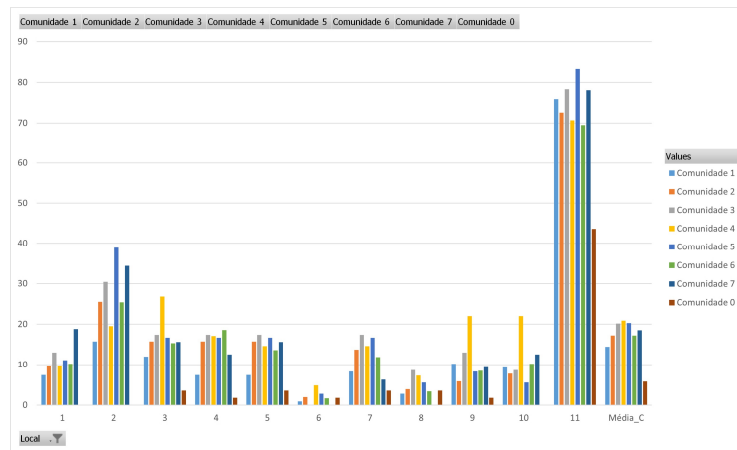


Figura 5.13: Resultados obtidos sobre a influência (Eixo Y) para os onze locais de eventos (Eixo X) e para as oito comunidades nas segundas-feiras.

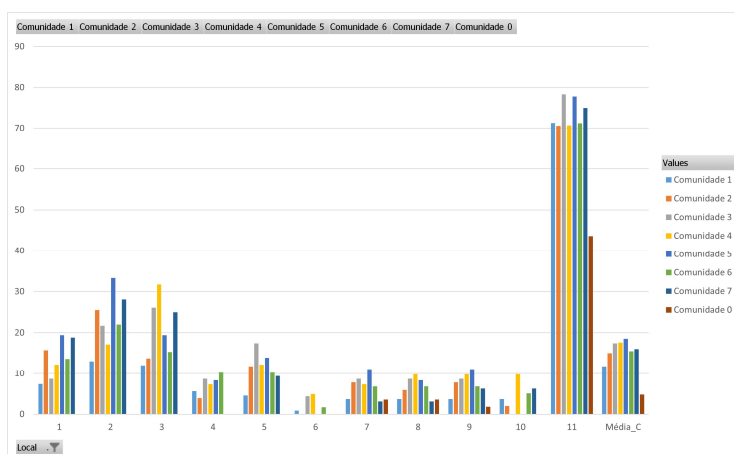


Figura 5.14: Resultados obtidos sobre a influência (Eixo Y) para os onze locais de eventos (Eixo X) e para as oito comunidades nas terças-feiras.

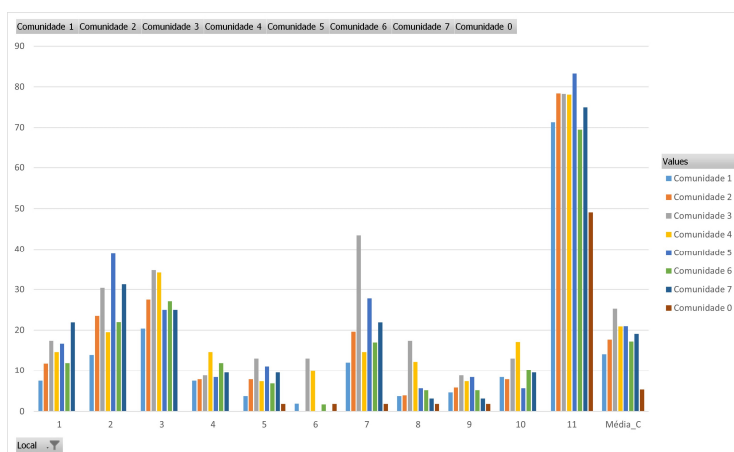


Figura 5.15: Resultados obtidos sobre a influência (Eixo Y) para os onze locais de eventos (Eixo X) e para as oito comunidades nas quartas-feiras.

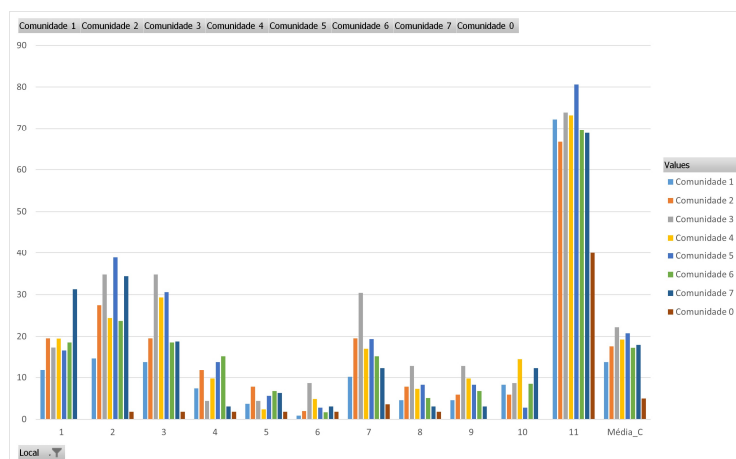


Figura 5.16: Resultados obtidos sobre a influência (Eixo Y) para os onze locais de eventos (Eixo X) e para as oito comunidades nas quintas-feiras.

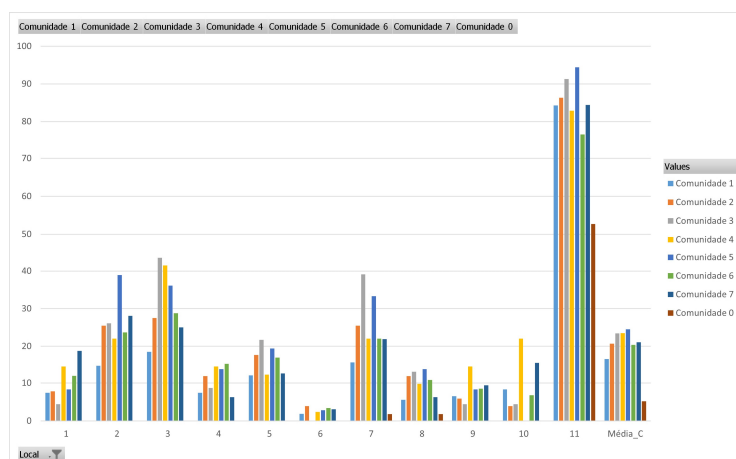


Figura 5.17: Resultados obtidos sobre a influência (Eixo Y) para os onze locais de eventos (Eixo X) e para as oito comunidades nas sextas-feiras.

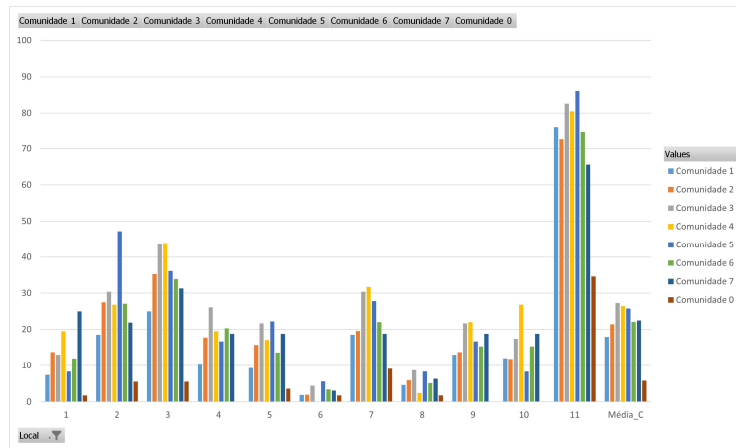


Figura 5.18: Resultados obtidos sobre a influência (Eixo Y) para os onze locais de eventos (Eixo X) e para as oito comunidades aos sábados.

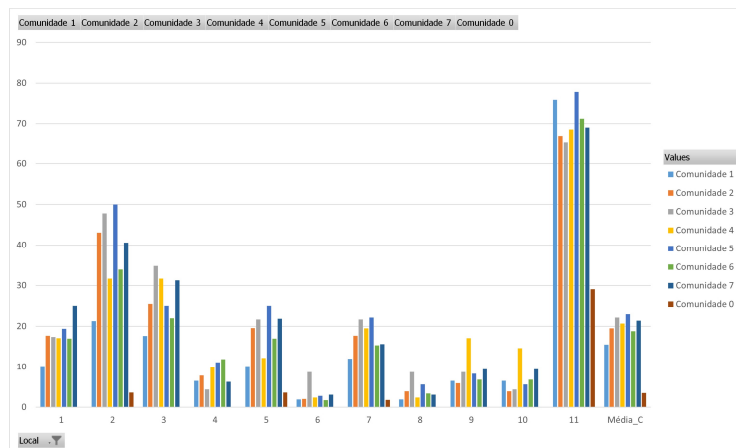


Figura 5.19: Resultados obtidos sobre a influência (Eixo Y) para os onze locais de eventos (Eixo X) e para as oito comunidades aos domingos.

Ao analisar os resultados mostrados nas figuras anteriores, é possível observar que a comunidade c_0 é influenciada em um nível muito menor que todas as demais comunidades independentemente do local e dia analisados. Esta característica provavelmente esta relacionada com o fato da comunidade c_0 ter sido formada a partir de um critério relacionado com a “ausência” de uma semântica específica referente à mobilidade de seus usuários.

Nota-se também que, para o local 11, os valores de influência obtidos para todas as comunidades e em todos os dias da semana são muitos superiores aos demais locais. Este fato é extremamente coerente ao se verificar que este local representa uma estação de trem muito movimentada.

Ao se excluir o local de evento 11 da análise, verifica-se que os locais 2 e 3 são os próximos em termos de influência sobre usuários em todos os dias da semana. Enquanto o local 2 influencia de maneira mais consistente nas Segundas, Terças, Quintas e Domingos, o local 3 influencia usuários com maior intensidade nas Quartas, Sextas e Sábados. É importante notar que, assim como o local de evento 11, o local de evento 3 também se encontra em uma estação de trem.

Em relação às comunidades, as comunidades c_3 , c_4 e c_5 foram as mais influenciadas pelos eventos em geral. Este fato é comprovado a partir da análise das médias da influência que cada evento exerceu sobre as comunidades. A comunidade c_3 foi a mais influenciada nas Quartas, Quintas, Sábados e Domingos, enquanto que a comunidade c_4 foi a mais influenciada nas Segundas e a comunidade c_5 nas Terças e Sextas.

A comunidade c_7 sofreu as maiores influências para o local de evento 1 em todos os dias da semana, com exceção da Terça. Este dado contribui para a indicação que o local de evento 1 é de extrema relevância para a comunidade c_7 . Por outro lado, o local de evento 6 é o que possui a menor relevância geral quando analisadas todas as comunidades, sendo sua média de influência inferior a 4, para todos os dias da semana.

5.2.4 Aplicação dos Resultados de Influência - Estudo de Caso 1

Nesta subseção, utiliza-se um caso hipotético no qual se poderia aplicar os resultados obtidos com a abordagem proposta, na qual foram identificados os níveis de influência de eventos no centro urbano de Pequim.

Neste cenário hipotético, descreve-se a intenção que o governo municipal de Pequim possui em implementar uma política pública de saúde através de campanhas de vacinação voltadas para grupos ou comunidades classificadas como de risco para a contração de uma determinada doença. Dado uma base de dados disponível que relaciona os custos horários específicos por dia da semana de locação ou reserva de uma local para a realização destas campanhas de vacinação, e segundo os resultados da análise de influência de eventos que podem ocorrer nestes locais, é possível reali-

zar a modelagem de um problema de otimização para que se possa obter o máximo de eficiência na realização destas campanhas de vacinação para uma comunidade k , dado uma restrição de orçamento.

Supondo que o alcance desta campanha de vacinação é diretamente relacionado ao número de horas de vacinação em cada local i e para cada dia da semana j , ponderado pela influência desse par i, j , realiza-se a seguinte formulação:

$$f(k) = \max \left(\sum_{i=1}^{11} \sum_{j=\text{segunda}}^{\text{domingo}} I_{ij}(k) \cdot y_{ij} \right)$$

sujeito a

$$\sum_{i=1}^{11} \sum_{j=\text{segunda}}^{\text{domingo}} x_{ij} \cdot y_{ij} \leq \text{budget}$$

$$y_{ij} \leq 16$$

$$x_{ij}, y_{ij} \geq 0$$

onde $I_{ij}(k)$ é a influência do local i no dia j , cujos valores podem ser obtidos nas Figuras 5.13, 5.14, 5.15, 5.16, 5.17, 5.18 e 5.19. y_{ij} refere-se à duração, em horas, da ação no local i no dia j . Os valores dessa variável foram limitados a um máximo de 16 horas por dia por cada local. A variável x_{ij} refere-se ao custo por hora do evento de vacinação no local i no dia j e *budget* é o orçamento disponível para essa ação.

De maneira a exemplificar a aplicação da metodologia neste problema hipotético, serão utilizados valores hipotéticos atribuídos ao problema de otimização definido acima considerando-se apenas o local 1, onde os custos e o orçamento estão representados em milhares de Yuans para a comunidade c_1 . Para o orçamento, definiu-se 100 mil Yuans. Os demais valores de influência e custos para o local 1 por dia da semana são apresentados na Tabela 5.2.

#	Segunda	Terça	Quarta	Quinta	Sexta	Sábado	Domingo
Influência	7,4	7,4	7,4	7,4	12	7,4	10,2
Custos	1,2	1	1	1	1,2	1,5	2

Tabela 5.2: Influência e Custos das ações por dia da semana.

De forma a desenvolver e encontrar uma solução para o problema de otimização formulado acima, é utilizado o método Simplex. O método Simplex é um método desenvolvido para a resolução de problemas de programação linear e em problemas de otimização combinatória.

Com o auxílio do método Simplex foram obtidos os valores ótimos para o número de horas de vacinação no local de evento 1, conforme o dia da semana j . Os resultados obtidos acerca da distribuição de horas de campanha são apresentados na Tabela 5.3.

Esses valores indicam que, para a distribuição de custo horário do local de evento 1 por dia da semana do exemplo, para a melhor utilização dos recursos deveriam ser realizadas ações durante 16h de segunda a sexta, 1h no sábado e 6h no domingo. Esse resultado gera a maior influência na comunidade c_1 dado o orçamento limite de 100 mil Yuans.

Segunda	Terça	Quarta	Quinta	Sexta	Sábado	Domingo
16	16	16	16	16	1	6

Tabela 5.3: Distribuição ótima de horas por dia da semana obtida através do modelo de otimização proposto.

5.3 Experimento 2

Neste segundo experimento, é apresentado o desenvolvimento dos passos e etapas necessários para a aplicação da metodologia proposta de forma a melhor explorar o potencial da mesma. Assim como no primeiro experimento detalhado na Seção 5.2, foi utilizado o *dataset* GEOLIFE [3], com suas características apresentadas na Seção 5.1.1.

5.3.1 Análise e Tratamento dos Dados

Como primeiro passo neste experimento, após a análise da distribuição temporal dos registros do *dataset* e com base na análise da Figura 5.3, foi realizada a filtragem dos dados, sendo utilizados apenas os dados no período entre os anos de 2008 e 2009. Esta filtragem foi realizada pois este é o intervalo com a maior densidade de dados existentes no *dataset*.

Após esta primeira filtragem dos dados, realizou-se a remoção de dados vazios e anormais. Então foram feitos o tratamento e a remoção de dados considerados duplicados. Foram considerados dados duplicados todos os registros que, para um dado usuário u_o , durante um dado intervalo de tempo, possuísem mais de um registro. O intervalo de tempo considerado foi de um minuto.

Para que esta remoção de dados duplicados fosse realizada sem causar distorções nos registros a serem utilizados no experimento, foi realizado o procedimento descrito

a seguir para a remoção destes dados.

Para cada intervalo de tempo de um minuto do *dataset*, para cada usuário u_o , calculou-se a localização média dos elementos de cada intervalo e se substituiu os registros duplicados pelo novo registro composto pela médias destes registros. A Figura 5.20 ilustra o processo de remoção de registros duplicados.

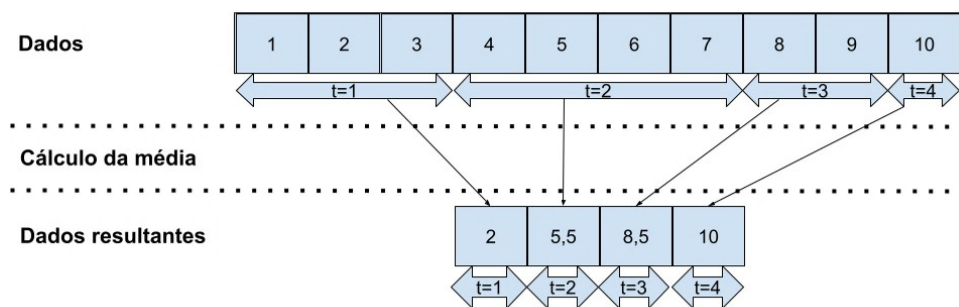


Figura 5.20: Ilustração do processo de remoção de registros duplicados dentro de um intervalo de tempo.

Os dados de mobilidade dos usuários, que estão disponíveis no *dataset* GEOLIFE, foram coletados durante intervalos de 1 a 5 segundos ou a cada 5 a 10 metros. Desta forma, no pior cenário do tratamento realizado para a remoção de dados duplicados, existiria uma distorção máxima de 300 metros. Este valor foi obtido considerando que para o pior caso, existiria para um dado usuário para um dado minuto, 60 registros com distância entre cada um deles de 10 metros, o que resultaria na distância máxima entre o primeiro registro e o último de 600 metros e, assim, após realizar a média entre os registros se obtém o valor de 300 metros.

Por outro lado, no melhor cenário, esta pequena distorção seria de 0 metros. Este valor pode ser obtido em três cenários distintos: a) no caso de haver apenas um registro naquele minuto para aquele usuário, b) se para aquele usuário, todos os registros naquele minuto estavam no mesmo local ou c) se dentro de um mesmo minuto aquele usuário se deslocou e voltou para o mesmo ponto. No caso médio,

a distorção gerada por esse tratamento será pequena de forma que as vantagens se sobrepõem às desvantagens da utilização deste tratamento.

Com este tratamento e esta filtragem dos dados, foi possível reduzir a quantidade de dados de cerca de 25 milhões de registros para pouco mais de 1 milhão de registros. Esta redução acarreta na redução do poder e dos recursos computacionais necessários para a execução do experimento, além de contribuir para que cada registro possua o mesmo peso em relação ao tempo durante os cálculos realizados.

5.3.2 Classificação dos usuários em comunidades

A classificação dos usuários em comunidades é realizada utilizando atributos disponíveis e fontes de dados, ou obtidos através de análise e mineração de dados de forma similar ao realizado no experimento da Seção 5.2.

Utilizando semântica distinta do primeiro experimento, neste experimento é realizada a classificação dos usuários utilizando como critério os atributos sobre o histórico de viagens dos usuários. Após a análise dos dados presentes no *dataset*, observou-se que um grupo de usuários possuía dados de localização coletados em regiões fora da área mais observada no *dataset*, a cidade de Pequim. Cerca de 46 usuários tiveram essa característica observada, enquanto os demais 74 usuários não possuem dados com estas características.

Para realizar esta separação de usuários viajantes dos usuários não viajantes, foi analisado as coordenadas que representava as fronteiras da cidade de Pequim, e monitorou-se os usuários que se deslocaram para fora desta fronteira. Cada usuário que deslocou-se para fora desta fronteira foi marcado como viajante e assim inserido na comunidade viajantes. Os demais usuários que não se deslocaram para fora da fronteira definida pela região da grande Pequim, foi inserido na comunidade dos não viajantes.

Assim sendo, utilizou-se este critério para a classificação dos usuários em comunidades formadas com a semântica de viajantes e não viajantes, sendo a comunidade c_1 a comunidade de viajantes e a comunidade c_2 a comunidade de não viajantes. Foi criada também uma terceira comunidade c_3 , comunidade de controle, com todos os usuários de forma a facilitar possíveis comparações, uma vez que as comunidades criadas possuem diferentes tamanhos.

Ao contrário do primeiro experimento, neste experimento não foram utilizados algoritmos de clusterização para a formação de comunidades. Realizou-se a classifi-

cação dos usuários em comunidades utilizando a seleção manual de atributos, neste caso, utilizando o critério “viajante”.

5.3.3 Definição de eventos

Neste experimento, serão utilizadas as mesma definições de eventos utilizados no experimento da Seção 5.2, apresentados a seguir.

Os locais selecionados para a ocorrência dos eventos são listados abaixo:

1. **Via de acesso ao aeroporto internacional de Pequim** (40.06519, 116.5882);
2. **5th Ring Road, próximo ao museu de ciência e tecnologia da China** (40.02228, 116.41758);
3. **Estação de trem “Beijing West”** (39.89716, 116.32107);
4. **Centro de Tênis de Pequim** (39.85218, 116.4138);
5. **Universidade de Tecnologia de Pequim** (39.8726, 116.48573);
6. **Clube Internacional de Golfe de Pequim** (39.959684, 116.234717);
7. **3 Qing Long Qiao Hao Jie** (40.004911, 116.268255);
8. **Clube de Golfe de Pequim Yanxi** (39.946325, 116.215314);
9. **Laoshan Velodrome** (39.914265, 116.211974);
10. **Parque ”Mundo”de Pequim** (39.812638, 116.279310);
11. **Estação de trem ”Qinghuayuan”** (39.982417, 116.339780).

Para cada um desses onze locais de eventos apresentados acima, são definidos sete cenários em que cada evento será analisado. Estes cenários são listados a seguir:

1. Eventos ocorrendo às Segundas-feiras;
2. Eventos ocorrendo às Terças-feiras;
3. Eventos ocorrendo às Quartas-feiras;
4. Eventos ocorrendo às Quintas-feiras;
5. Eventos ocorrendo às Sextas-feiras;
6. Eventos ocorrendo aos Sábados;

7. Eventos ocorrendo aos Domingos;

O Fator de Influência para cada evento neste experimento foi definido como $IF = 1$ devido ao não conhecimento prévio da intensidade, ao longo do tempo, dos eventos selecionados. Além disso, optou-se pela utilização de $IF = 1$ com o objetivo de realizar a comparação de eventos de maneira linear ao longo do tempo de forma a não adicionar pesos, maiores ou menores, a intervalos específicos do tempo de duração destes eventos.

5.3.4 Apresentação dos Resultados

Para cada um dos onze locais de eventos, calculou-se a influência para cada um deles para cada semana dentro de todo período do *dataset*. Com a combinação dos onze locais de eventos com todos os sete cenários listados anteriormente, foi realizada a análise de ao todo 77 possibilidades de eventos.

Cada nome de evento é composto pela letra "E" seguida pelo identificador do local de evento, seguido por três letras para a identificação do dia da semana. A título de exemplo, um evento com o nome "E01 - MON" representa um evento que ocorre no local de evento 01 (Via de acesso ao aeroporto internacional de Pequim) no dia da semana "Segunda-feira".

Por simplicidade, não será apresentada a totalidade dos resultados obtidos, sendo selecionados alguns resultados para a discussão. Selecionou-se os resultados dos usuários mais influenciados por eventos para a apresentação detalhada de seus resultados. Na Tabela 5.4, é mostrada a lista dos usuários (com seus números de identificação) mais influenciados por eventos e os eventos que exerceram influência sobre usuários com mais intensidade, quando comparado com os demais.

Tabela 5.4: Lista dos usuários mais influenciados (à esquerda) e lista dos eventos que exerceram influência com maior intensidade (à direita).

Usuários mais influenciados	Comunidade	Eventos mais influenciadores
128	C2	E11 - SAT
140	C2	E11 - SUN
85	C2	E11 - MON
17	C2	E11 - TUE
144	C2	E11 - WED
84	C2	E11 - THU
14	C1	E11 - FRI
167	C2	E02 - MON
115	C2	E03 - SAT
155	C1	E03 - SUN

Para apresentação dos resultados de influência obtidos dos usuários, foram selecionados da Tabela 5.4 os usuários 128 e 140, para a comunidade dos que não viajaram para locais fora da região da cidade de Pequim, e os usuários 14 e 155, para a comunidade de usuários que viajaram para fora da cidade de Pequim. Para estes quatro usuários, serão apresentados os resultados para os eventos "E11 - SAT" e "E11 - SUN".

Esses quatro usuários foram escolhidos devido ao fato desses terem sido os mais influenciados por ambos os eventos ao longo do tempo. Já os eventos, por eles terem influenciado usuários com maior intensidade que os demais.

Na Figura 5.21 (a), (b), (c) e (d) são apresentadas as influências para cada instante de tempo para os usuários 128, 140, 14 e 155 durante o tempo de monitoramento do evento 'E11 - SAT', respectivamente.

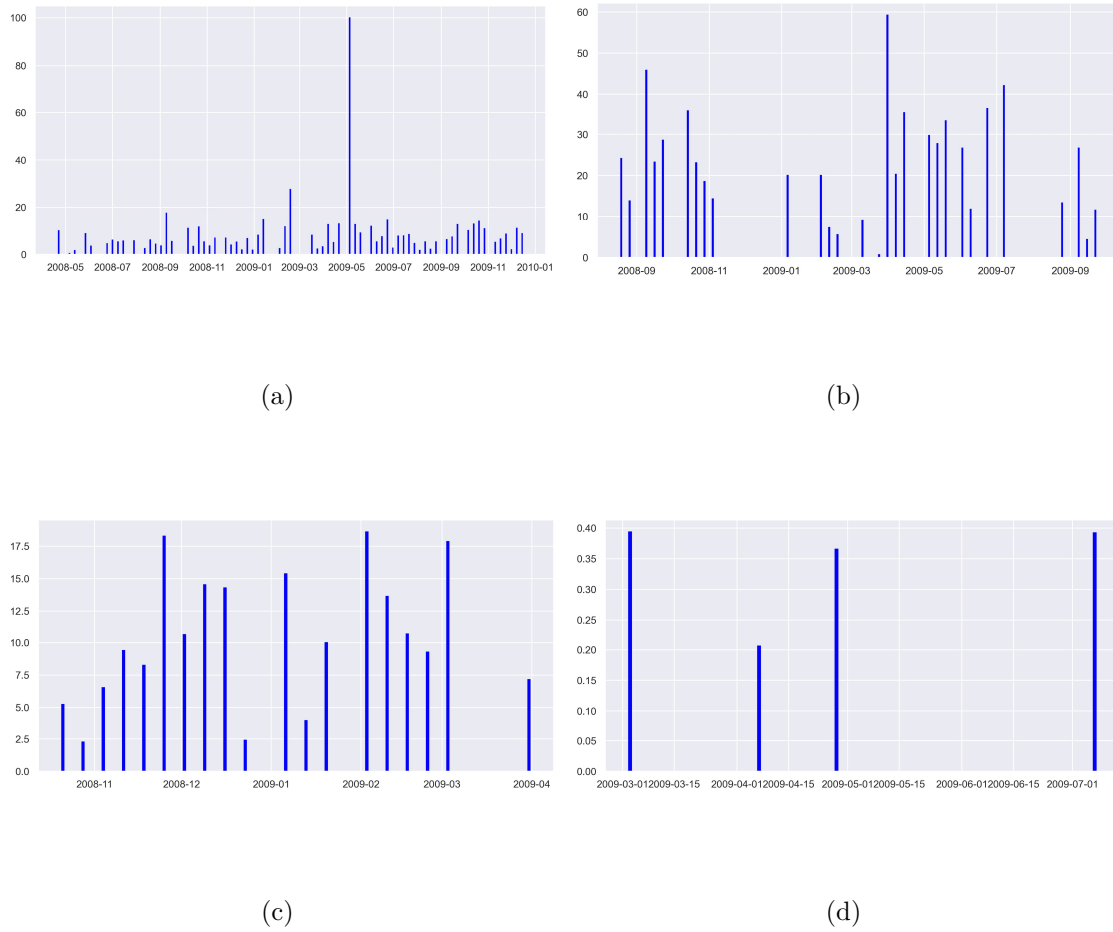


Figura 5.21: Influência para cada instante de tempo de monitoramento do evento 'E11 - SAT' para os usuários 128 (a), 140 (b), 14 (c) e 155 (d), respectivamente.

Na Figura 5.22 (a), (b), (c) e (d) são apresentadas as influências para cada instante de tempo para os usuários 128, 140, 14 e 155 durante o tempo de monitoramento do evento 'E11 - SUN', respectivamente.

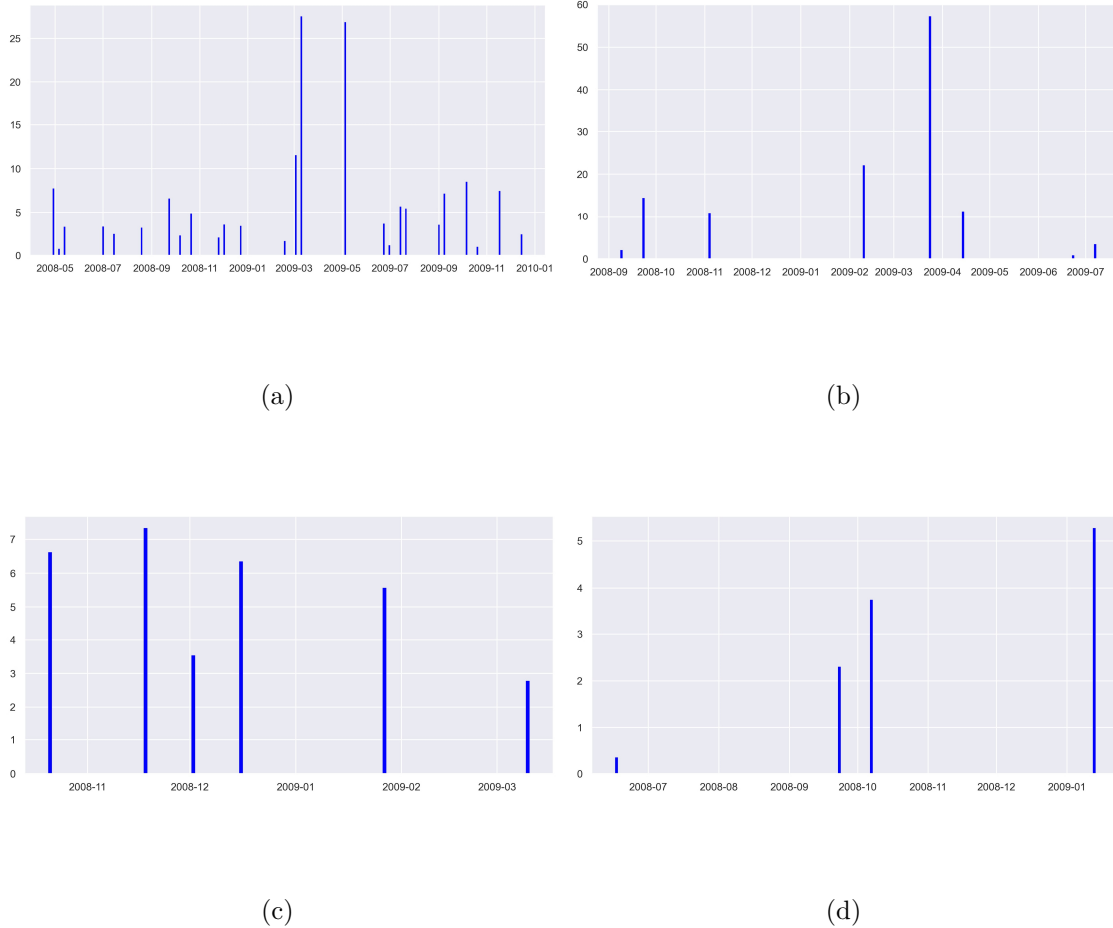


Figura 5.22: Influência para cada instante de tempo de monitoramento do evento 'E11 - SUN' para os usuários 128 (a), 140 (b), 14 (c) e 155 (d), respectivamente.

Como pode ser visto nas influências apresentadas nas Figuras 5.21 (a), (b), (c) e (d) e nas Figuras 5.22 (a), (b), (c) e (d), elas são bastante contrastantes entre si. Para estes usuários, pode-se notar, ao se comparar os valores de influência para os eventos 'E11 - SAT' e 'E11 - SUN', que o primeiro evento é mais relevante que o último, com exceção apenas para o usuário 155.

Após apresentar os resultados de alguns usuários e eventos, são apresentadas na Figura 5.23 (a) e (b) as Influência Média dos eventos 'E11 - SAT' e 'E11 - SUN' para todos os usuários de todas as comunidades. Nestas figuras, o eixo X refere-se ao identificador do usuário e o eixo Y refere-se à Influência Média obtida do evento

sobre cada usuário.

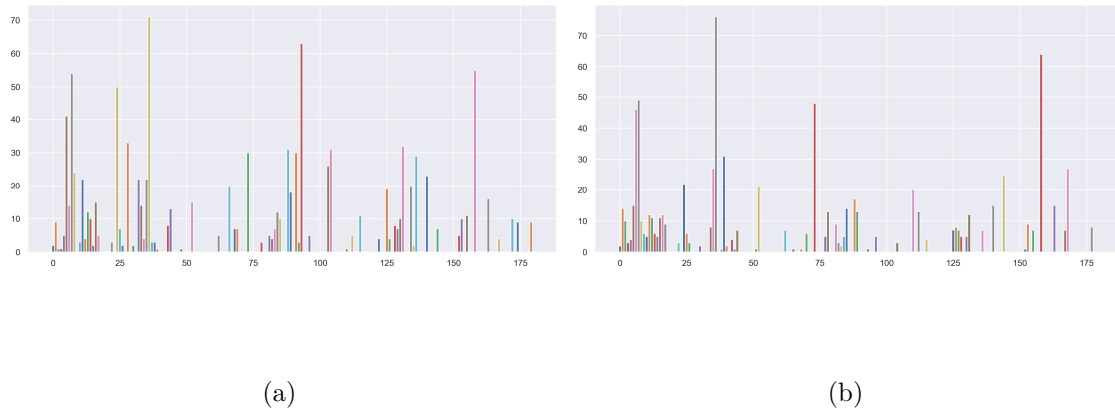


Figura 5.23: Influência Média para o evento 'E11 - SAT' para todos usuários de ambas as comunidades (a), e Influência Média do evento 'E11 - SUN' para todos usuários de ambas as comunidades (b).

Os usuários 128, 140, 14 e 155 foram influenciados pelo evento 'E11 - SAT' em diferentes graus de influência, sendo os valores de influência obtidos iguais a 8.06, 23.03, 10.44, e 11.59, respectivamente. Para o evento 'E11 - SUN', foram obtidos como níveis de influência deste evento nesses usuários os valores de 5.98, 15.15, 5.34 e 7.08, respectivamente.

Já ao se analisar a influência de eventos sobre comunidades, a comparação é realizada de duas formas distintas: a primeira, comparando eventos que ocorrem no mesmo local, de forma a analisar a variação da influência ao longo da semana, facilitando assim a comparação dia a dia; e a segunda, realizando a comparação de eventos que ocorrem em diferentes locais, porém em um mesmo dia da semana e assim, desta forma, facilitando a comparação da influência de diferentes locais dado um mesmo período de ocorrência deste evento.

Estas comparações são utilizadas com o objetivo de identificar similaridades e diferenças no comportamento de eventos através da análise da influência gerada sobre comunidades. Na Figura 5.24, é apresentado um gráfico para cada um dos onze locais de eventos. Cada gráfico representa a influência daquele evento para cada uma das três comunidades estudadas, em cada dia da semana.

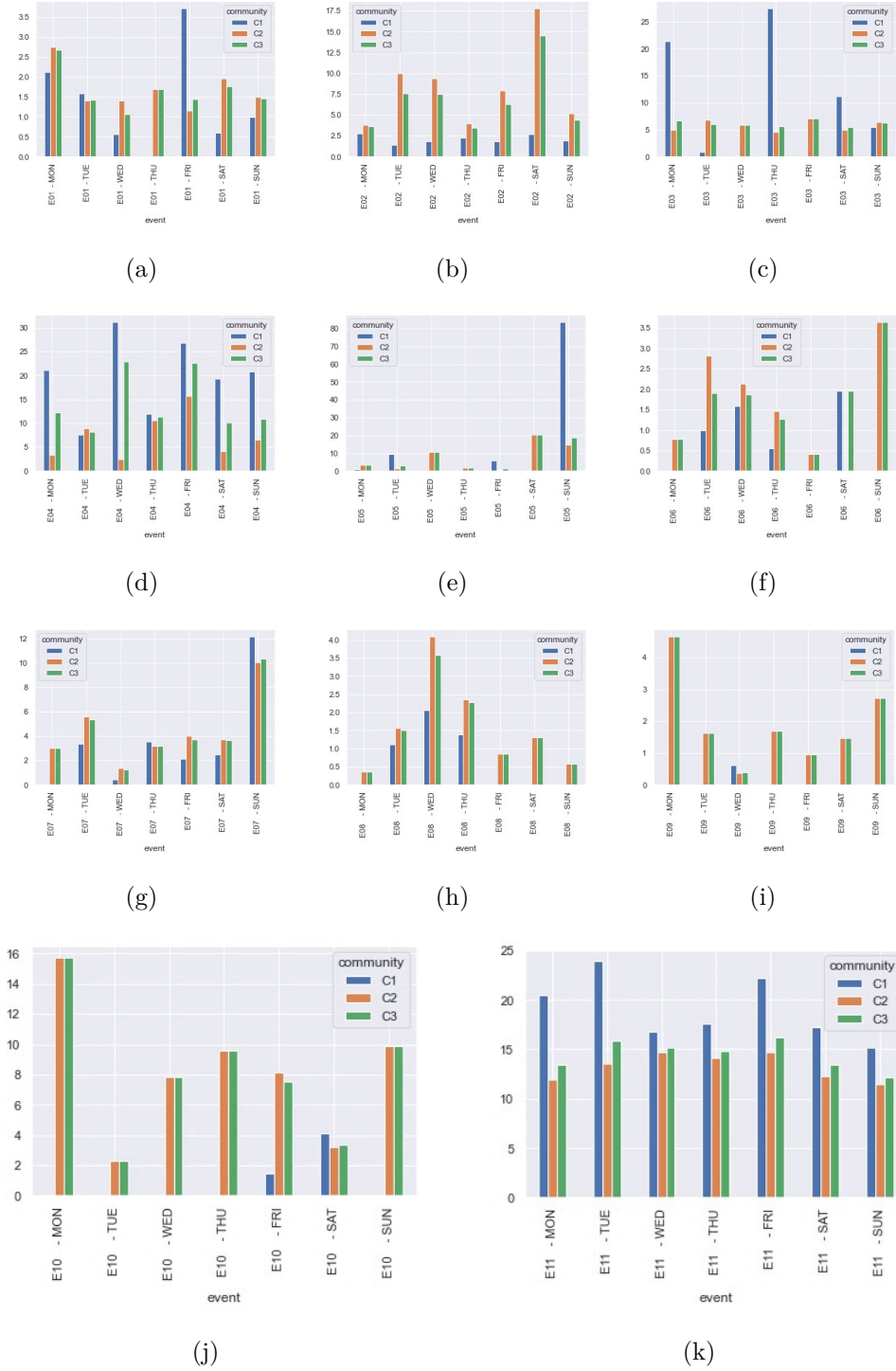


Figura 5.24: Influência média do evento "E1" (a), "E2" (b), "E3" (c), "E4" (d), "E5" (e), "E6" (f), "E7" (g), "E8" (h), "E9" (i), "E10" (j) e "E11" (k) em cada dia da semana para todos comunidades.

Como pode ser visto na Figura 5.24, alguns eventos influenciam com maior intensidade algumas comunidades quando comparadas com outras. Os eventos "E2", "E6", "E8", "E9" e "E10" afetam com mais intensidade a comunidade "C2" do

que a comunidade "C1". Já os eventos "E4" e "E11" afetam mais a comunidade "C1" quando comparado com a comunidade "C2". Os demais eventos, "E1", "E3", "E5" e "E7" afetam mais a comunidade "C1" em alguns dias da semana e em outros a comunidade "C2", de maneira alternada. Com base no exposto, é possível afirmar acerca da existência de indícios suficientes para dizer que existem locais de eventos mais relevantes que outros em termos de potencial de influenciar determinados usuários e comunidades.

Na Figura 5.25, são apresentados gráficos para cada dia da semana mostrando todos os onze eventos. Cada gráfico apresenta a influência obtida de cada evento para cada uma das três comunidades.

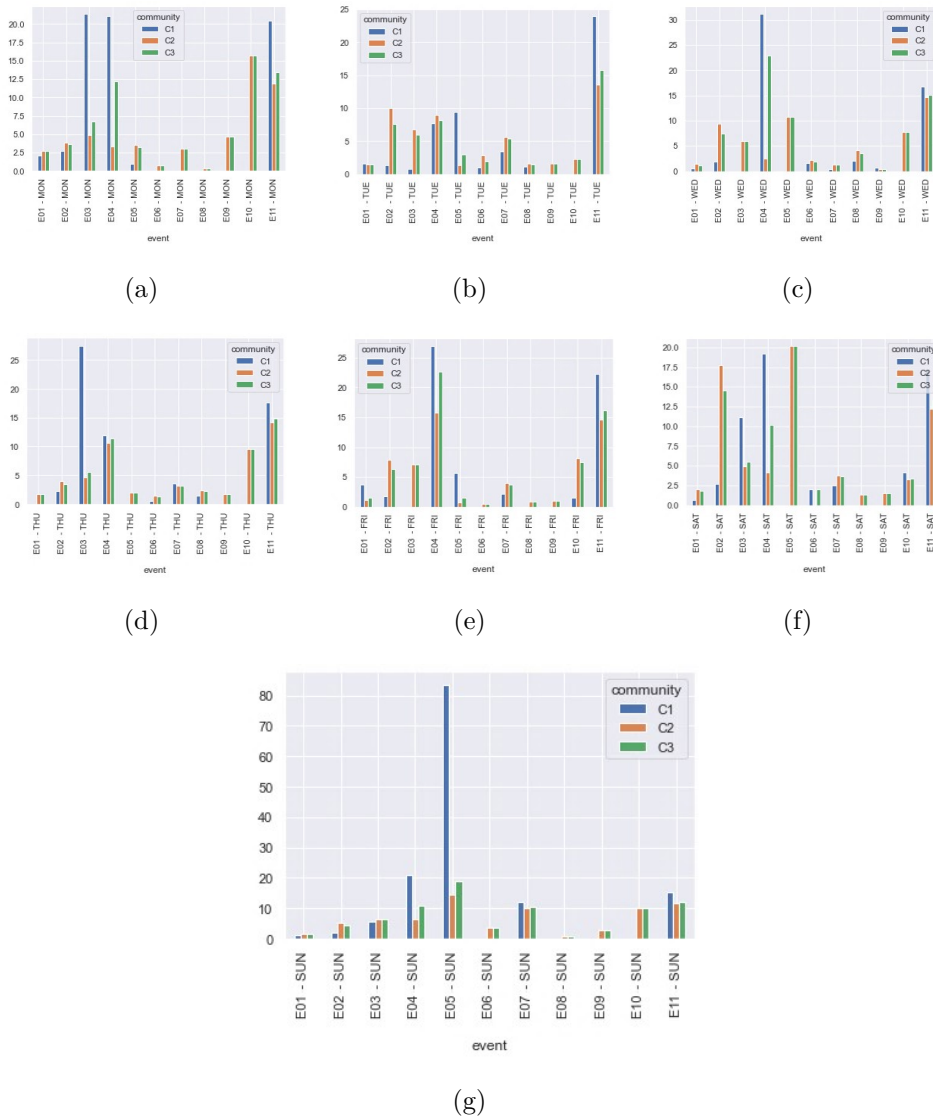


Figura 5.25: Influência média de todos onze eventos na **Segunda** (a), na **Terça** (b), na **Quarta** (c), na **Quinta** (d), na **Sexta** (e), no **Sábado** (f) e no **Domingo** (g).

Como pode ser visto na Figura 5.25, eventos afetam comunidades de formas diferentes mesmo durante os mesmos dias da semana. Esta característica se deve provavelmente ao fato da existência de diferentes características de comportamento entre comunidades ao longo da semana.

Por todo o exposto, conclui-se que comunidades e locais de eventos, e por consequência comunidades e eventos, são de alguma forma relacionados. Esta relação entre comunidades e eventos é analisada neste trabalho através do conceito de influência.

5.3.5 Aplicação dos Resultados de Influência

Existem diversos casos em que é possível aplicar a metodologia proposta, nas mais distintas áreas de conhecimento, como saúde, segurança, educação, mobilidade urbana, transporte, *marketing* etc. Nesta seção são apresentados dois estudos de casos para ilustrar como a metodologia proposta pode ser aplicada em dois diferentes problemas: (a) uma campanha de vacinação e (b) uma campanha de *marketing*.

5.3.5.1 Estudo de Caso 1 - Campanha de Vacinação

O primeiro estudo de caso é a intenção do Governo da cidade de Pequim de realizar a implantação de uma campanha de vacinação voltada para um público alvo com alto risco de contaminação de uma determinada doença. O governo da cidade de Pequim obteve evidência suficiente de que pessoas que realizaram viagens para fora da cidade de Pequim e para o exterior são consideravelmente mais prováveis de contrair esta determinada doença. Desta forma, as autoridades de Pequim iniciaram os preparativos para a realização de uma campanha de vacinação focada principalmente no grupo de pessoas que tenha realizado viagens para fora da cidade de Pequim, em outras palavras, a comunidade "C1".

Se o custo por hora por dia da semana de utilização de um dado local que pode receber as instalações de uma campanha de vacinação é conhecida, é possível formular um problema de otimização para a obtenção da melhor distribuição de horas por cada local de evento por dia da semana para uma dada restrição orçamentária.

Assumindo que a eficiência de uma campanha de vacinação é diretamente relacionada ao número de horas gastas em cada local de evento i imunizando usuários em cada dia de semana j , ponderado pelo nível de influência para o par i, j , e também considerando os onze locais de eventos e outros dados apresentados nas seções

anteriores, tem-se a seguinte formulação do problema:

$$f(k) = \max(\sum_{i=1}^{11} \sum_{j=1}^7 I_{ij}(k) \cdot y_{ij})$$

Sujeito a

$$\sum_{i=1}^{11} \sum_{j=1}^7 x_{ij} \cdot y_{ij} \leq \text{budget}$$

$$\frac{y_{ij}}{n_teams} \leq 8$$

$$x_{ij}, y_{ij} \geq 0$$

onde, $I_{ij}(k)$ é o nível de influência do local de evento i no dia da semana j , cujos valores podem ser obtidos na Figura 5.25 (a), (b), (c), (d), (e), (f) e (g). A variável n_teams refere-se ao número de equipes disponíveis para atuar nesta campanha de vacinação. y_{ij} refere-se a duração, em horas, da campanha no local de evento i no dia da semana j , limitado a um máximo de oito horas por dia por equipes de vacinação em cada local de evento. x_{ij} refere-se ao custo por hora de uso do local de evento i no dia da semana j e $budget$ que refere-se a limitação de orçamento para esta campanha de vacinação.

Por simplicidade, neste estudo de caso os valores atribuídos a x_{ij} serão definidos como 1 (mil dólares) para todos os locais de evento por hora. Foi definido também o valor para a variável $budget$ sendo igual a 250 (milhares de dólares). Foi definido também o número de equipes disponíveis para a realização desta campanha de vacinação (n_teams) igual a 5. Estes times podem trabalhar no mesmo local de evento ou em locais diferentes, sendo sua atribuição a um local dependente exclusivamente da eficiência de cada local de evento em um dado dia da semana, e sendo desconsiderados possíveis tempos de deslocamento entre localidades com o objetivo de simplificar a análise do estudo de caso. Os níveis de influência para a comunidade C2 em todos os locais de evento e em cada dia da semana é conhecido e apresentado na Tabela 5.5.

Como resultado da aplicação do método Simplex para a formulação deste problema, o número ótimo de horas por dia da semana considerando todos os locais de eventos para a comunidade C2 é apresentado na Tabela 5.6.

Como pode ser visto na Tabela 5.6 alguns dos locais de eventos foram mais eficientes do que outros e, devido a este fato, eles foram selecionados para hospedar

Tabela 5.5: Valores de Influência obtidos para cada dia da semana para todos os locais de eventos para a comunidade “C2” para a campanha de vacinação modelada neste estudo de caso.

#Influência	Segunda	Terça	Quarta	Quinta	Sexta	Sábado	Domingo
Local 1	2,74	1,39	1,4	1,69	1,15	1,95	1,49
Local 2	3,83	9,96	9,4	4	7,91	17,74	5,15
Local 3	4,89	6,77	5,87	4,57	7,09	4,96	6,36
Local 4	3,36	8,95	2,46	10,54	15,74	4,17	6,45
Local 5	3,59	1,37	10,69	1,97	0,76	20,14	14,54
Local 6	0,77	2,82	2,12	1,46	0,41	0	3,63
Local 7	3	5,61	1,34	3,18	4,01	3,71	10,02
Local 8	0,37	1,56	4,08	2,36	0,85	1,3	0,58
Local 9	4,63	1,61	0,36	1,68	0,94	1,47	2,7
Local 10	15,7	2,27	7,81	9,58	8,12	3,25	9,89
Local 11	11,93	13,52	14,72	14,15	14,65	12,25	11,49

Tabela 5.6: Distribuição ótima de horas de campanha vacinação por local de evento e por dia da semana com restrição orçamentária.

#Horas	Segunda	Terça	Quarta	Quinta	Sexta	Sábado	Domingo	Soma
Local 1	0	0	0	0	0	0	0	0
Local 2	0	16	2	0	0	16	0	34
Local 3	0	0	0	0	0	0	0	0
Local 4	0	0	0	16	16	0	0	32
Local 5	0	0	16	0	0	16	16	48
Local 6	0	0	0	0	0	0	0	0
Local 7	0	0	0	0	0	0	8	8
Local 8	0	0	0	0	0	0	0	0
Local 9	0	0	0	0	0	0	0	0
Local 10	16	0	0	8	0	0	0	24
Local 11	16	16	16	16	16	8	16	104
Soma	32	32	34	40	32	40	40	-

esta campanha hipotética de vacinação. Os locais de evento 01, 03, 06, 08 e 09 não foram selecionados nem ao menos uma vez. Os dias da semana Quinta-feira, Sábado e Domingo foram os dias da semana mais eficientes. Ao analisar estes resultados, é possível observar que, se for aumentado o valor do orçamento disponível, é possível aumentar o número de horas nesta campanha de vacinação nas Segundas-feiras, Terças-feiras, Quartas-feiras e Sextas-feiras, até que ele atinja o limite de número de times disponíveis.

Neste estudo de caso, foi apresentada a aplicação da metodologia proposta e a forma em que ela poderia ser utilizada para auxiliar no aumento de eficiência ao aplicá-la neste e em diversos outros cenários. Foi demonstrado que é possível realizar a redução de recursos e alcançar os mesmo resultados. No caso apresentado, foi possível realizar a redução da utilização de locais disponíveis para hospedar esta campanha de vacinação de um total de 11 locais de eventos para apenas 6 locais.

5.3.5.2 Estudo de Caso 2 - Campanha de *Marketing*

O segundo estudo de caso é um campanha de hipotética de *marketing* para uma grande companhia de tecnologia, Big Tech S.A., utilizando *outdoors* digitais espalhados pela cidade de Pequim. Diferentemente da campanha de vacinação, a empresa hipotética de publicidade Marketing S.A., contratada pela Big Tech S.A. para a realização de publicidade, possui orçamento virtualmente ilimitado para a realizar a publicidade desta campanha de *marketing* pela cidade de Pequim. Porém, o contrato realizado entre as duas companhias possui uma cláusula que especifica um alcance mínimo de usuários para esta campanha. Além disso, Big Tech S.A. deseja que o público alvo desta campanha seja formada por usuários que não têm o hábito de realizar viagens para fora da cidade de Pequim, uma vez que a grande maioria de seus clientes possuem este perfil, conforme pesquisas (hipotéticas) realizadas anteriormente.

Marketing S.A. possui painéis digitais espalhados pela cidade de Pequim e conhece o custo de publicidade em cada um destes locais, devido aos custos de locação. Como o custo por hora por dia da semana da utilização de cada painel digital é conhecido, e sabendo que a Marketing S.A. é uma empresa com fins lucrativos e que ela deseja obter o maior lucro possível, pode-se realizar a formulação de um problema de otimização no qual se deseja obter a melhor distribuição ótima de horas desta campanha de *marketing* por local de painel digital por dia da semana dado o alcance de visualização deste painel por usuários e com restrição mínima de alcance destes usuários, de forma a minimizar os custos da campanha.

Considerando que o alcance de visualização de usuários desta campanha de *marketing* é diretamente relacionada com o número de horas gastas no local de evento, ou neste caso, local do painel digital, i , onde a peça publicitária para a Big Tech S.A. está sendo exibida em cada dia da semana j , ponderado pelo nível de influência do par i, j , e considerando todos os onze locais onde a Marketing S.A. possui painéis digitais pela cidade de Pequim, é possível realizar a seguinte formulação para o problema de otimização:

$$f(k) = \min(\sum_{i=1}^{11} \sum_{j=1}^7 x_{ij}(k) \cdot y_{ij})$$

Sujeito a

$$\sum_{i=1}^{11} \sum_{j=1}^7 I_{ij} \cdot y_{ij} \geq \text{inf_goal}$$

$$y_{ij} \leq 24$$

$$x_{ij}, y_{ij} \geq 0$$

onde $I_{ij}(k)$ é o nível de Influência do local i no dia da semana j , cujos valores podem ser obtidos da Figura 5.25 (a), (b), (c), (d), (e), (f) e (g). y_{ij} refere-se à duração, em horas, da campanha de *marketing* em cada local i e em cada dia da semana j , limitado ao máximo de 24 horas por local. x_{ij} refere-se ao custo total por hora de uso de cada painel digital do local i no dia da semana j e inf_goal é o alcance mínimo de usuários definidos em contrato entre as empresas para a prestação de serviços publicitários.

Neste caso em tela, os valores atribuídos a x_{ij} serão definidos em 1 (dólares) para todos os locais de evento por hora. Para a variável inf_goal é definido o valor 10000. Os níveis de Influência Média para a comunidade C1 em todos os locais de evento e para cada dia da semana são apresentados na Tabela 5.7.

Tabela 5.7: Valores de Influência Média obtidos para cada dia da semana para todos os locais de eventos para a comunidade C1 para a campanha publicitária.

#Influência	Segunda	Terça	Quarta	Quinta	Sexta	Sábado	Domingo
Local 1	2,11	1,57	0,55	n/a	3,7	0,59	0,99
Local 2	2,74	1,38	1,85	2,23	1,8	2,68	1,9
Local 3	21,35	0,78	n/a	27,41	n/a	11,18	5,49
Local 4	21,09	7,64	31,11	11,94	26,88	19,22	20,8
Local 5	0,99	9,43	n/a	n/a	5,71	n/a	83,5
Local 6	n/a	1	1,59	0,55	n/a	1,96	n/a
Local 7	n/a	3,35	0,41	3,55	2,15	2,46	12,12
Local 8	n/a	1,11	2,06	1,38	n/a	n/a	n/a
Local 9	n/a	n/a	0,62	n/a	n/a	n/a	n/a
Local 10	n/a	n/a	n/a	n/a	1,44	4,12	n/a
Local 11	20,44	23,89	16,79	17,57	22,22	17,26	15,13

Como resultado de aplicar o método simplex para esta formulação de problema, o número ótimo de horas por dia da semana considerando todos os locais com painéis digitais disponíveis para a comunidade C1 é apresentado na Tabela 5.8.

Como pode ser visto na Tabela 5.8, alguns locais foram mais eficientes que outros

Tabela 5.8: Distribuição ótima de horas da campanha de marketing por evento e por dia da semana dado uma restrição de influência de evento (alcance) na comunidade C1.

#Horas	Segunda	Terça	Quarta	Quinta	Sexta	Sábado	Domingo	Soma
Local 1	0	0	0	0	0	0	0	0
Local 2	0	0	0	0	0	0	0	0
Local 3	24	0	0	24	0	17,06	0	65,06
Local 4	24	0	24	24	24	24	24	144
Local 5	0	0	0	0	0	0	24	24
Local 6	0	0	0	0	0	0	0	0
Local 7	0	0	0	0	0	0	24	24
Local 8	0	0	0	0	0	0	0	0
Local 9	0	0	0	0	0	0	0	0
Local 10	0	0	0	0	0	0	0	0
Local 11	24	24	24	24	24	24	24	168
Soma	72	24	48	72	48	65,06	96	-

e, devido a isto, foram selecionados para hospedar a publicidade desta campanha hipotética de *marketing*.

Os locais de evento 01, 02, 06, 08, 09 e 10 não foram escolhidos nem ao menos uma vez devido à baixa eficiência que estes locais representaram. Domingo foi o dia da semana mais eficiente dentre todos os dias da semana, seguido por Segunda-feira e Quinta-feira. Locais 04 e 11 foram os locais mais eficientes para esta campanha publicitária. Ao realizar a análise destes resultados, é possível observar que, ao aumentar o alcance mínimo da campanha publicitária, é necessário aumentar o número de horas desta campanha de forma a atingir o novo nível da restrição, o que consequentemente elevaria os custos gerais da campanha de *marketing*.

A distribuição das peças publicitárias por local e dia da semana é apresentada na Tabela 5.8 e representa um custo total de 425.06\$ dólares por uma semana de campanha de marketing realizada pela Marketing S.A., com o lucro gerado pela campanha sendo igual à diferença da receita recebida pela campanha de publicidade com o custo total da realização da campanha.

Este estudo de caso demonstrou a utilização desta metodologia de forma a auxiliar no aumento de eficiência de eventos através da redução de custos, podendo ser utilizada em diversos cenários. Foi demonstrado também que é possível minimizar a utilização e o desperdício de recursos atingindo os mesmos resultados. No caso apresentado, foi possível obter a redução de possíveis locais para a realização da campanha de *marketing* de 11 potenciais locais para apenas 5, liberando recursos para a utilização em outras campanhas publicitárias.

5.4 Experimento 3

Neste terceiro experimento é apresentado o desenvolvimento das etapas necessárias e seus respectivos passos para melhor explorar o potencial da metodologia proposta.

Neste experimento, foi utilizado como fonte de dados o *dataset* CABSPOTTING [4], cujas características gerais foram apresentadas na Seção 5.1.2. Além disso, este experimento tem como objetivo comparar as características obtidas a partir da adoção de dois diferentes métodos de classificação de usuários em comunidades: (a) K-Means e (b) DBSCAN. Neste experimento é apresentada também a aplicação de uma função no tempo para a variação do Fator de Influência.

5.4.1 Análise e Tratamento dos Dados

O primeiro passo realizado neste experimento foi a análise espaço-temporal dos registros disponíveis neste *dataset*, conforme apresentado na Seção 5.1.2.

Após esta análise, observou-se que os usuários não possuíam padrões claros de mobilidade ao longo do tempo, ou seja, o comportamento dos usuários não apresenta um hábito claramente definido, de forma similar ao descrito por [29], com seus deslocamentos não ocorrendo de forma linear, como dito em [45]. Por este motivo, optou-se então por analisar também os dados disponíveis das corridas, as quais cada usuário realizou.

No *dataset* não são disponibilizadas as informações de corridas dos usuários, sendo disponíveis apenas as informações acerca da ocupação do usuário (*fare*), sendo "0" para táxi livre e "1" para táxi ocupado. A partir desta informação, foi realizado um tratamento dos registros de forma a identificar as corridas realizadas por cada usuário.

Foi definido como início de uma corrida ct todo registro no qual um táxi tx constasse como ocupado para um instante de tempo t_i e que constasse como livre no instante de tempo anterior t_{i-1} . De forma similar, foi definido como fim de uma corrida ct todo registro no qual um táxi tx constasse como ocupado para um instante de tempo t_i e que constasse como livre no instante de tempo posterior t_{i+1} .

A Figura 5.26(a) apresenta a distribuição espacial dos locais de início das corridas dos usuários, a Figura 5.26(b) apresenta a distribuição espacial dos locais de fim das

corridas e a Figura 5.26(c) apresenta a distribuição espacial de todos os locais de início e fim das corridas.

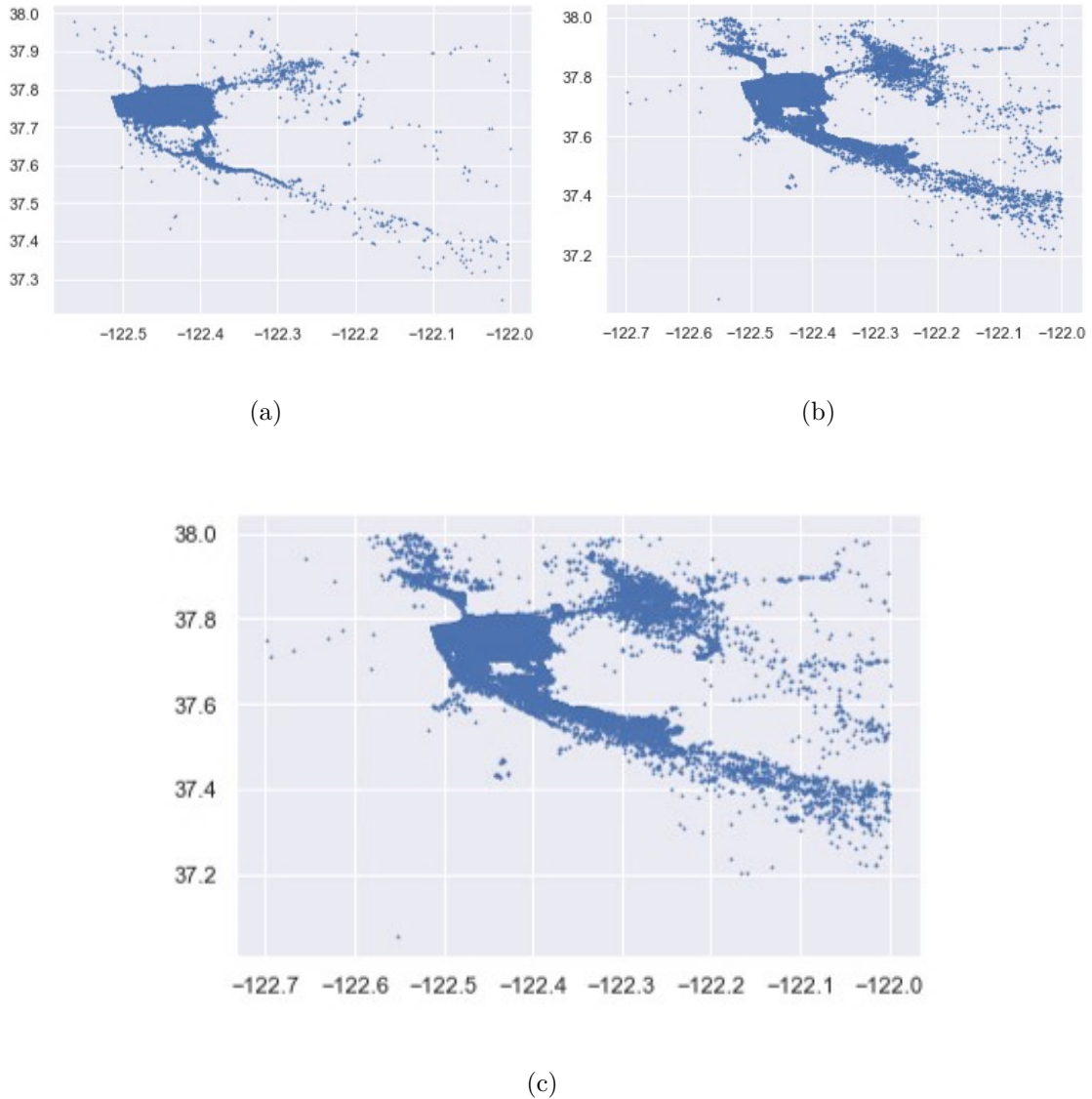


Figura 5.26: *Scatter plot* para os locais de início de corrida (a), para os locais de fim de corrida (b) e para os ambos locais de início e de fim de corrida de forma consolidada (c).

5.4.2 Classificação dos usuários em comunidades

A classificação dos usuários em comunidades deve, idealmente, ser realizada utilizando atributos que sejam relevantes para os usuários disponíveis, de forma a criar uma semântica em cada comunidade criada e uma relação de pertencimento destes usuários a cada uma das comunidades que ele pertença. Devido ao fato de usuários (táxis) não possuírem um comportamento de mobilidade claro e seus deslocamentos

não possuírem um comportamento habitual, decidiu-se realizar a classificação dos usuários utilizando como critério o padrão de suas viagens.

Para realizar a classificação dos usuários em comunidades, utilizou-se os dados dos locais de destino de corridas. Selecionou-se esta informação para utilizar como critério de criação de comunidades pois: (a) é relevante para a análise dos principais destinos de clientes dos táxis, (b) por cobrir uma maior área da região onde os registros do *dataset* estão disponíveis e (c) devido a uma análise prévia da distribuição dos táxis.

Os usuários (táxis) não possuem um padrão bem definido de mobilidade e, por isso, o comportamento de deslocamento não possui “memória”, ou seja, este deslocamento não ocorre de forma recorrente ou periódica, dificultando análises muito antecipadas em relação ao tempo. Com essa característica em evidência e visando não inserir vieses na análise dos dados deste experimento, para a classificação de usuários em comunidades realizou-se a seleção de dados de um dia da semana no período entre as 06:00 e as 12:00 e utilizou-se este subconjunto de dados para a classificação dos usuários em comunidades. Para realizar esta classificação, utilizou-se dois algoritmos de clusterização para fins de comparação: (a) K-Means e (b) DBSCAN.

Considerando as características dos dados e dos métodos de agrupamento utilizados e em função da distribuição temporal dos dados utilizados, espera-se que os usuários que serão classificados, pertençam a mais de uma comunidade, visto que estes usuários irão se deslocar da região de uma comunidade a outra ao longo do tempo.

5.4.2.1 Classificação dos usuários em comunidades utilizando K-Means

Para realizar a classificação dos usuários em comunidades utilizando o K-Means é necessária a definição de um número “k” de comunidades a priori da utilização do K-Means. Para realizar a escolha desse número, foi utilizado o método Elbow, cuja análise dos resultados apontou para a definição de $k = 10$. Os valores obtidos na análise do número “k” de comunidades utilizando o método Elbow são apresentados na Figura 5.27.

Após ter sido definido o valor de “k”, é realizada a classificação dos usuários propriamente dita. Após a execução do K-Means, os registros foram classificados conforme ilustrado na Figura 5.28.

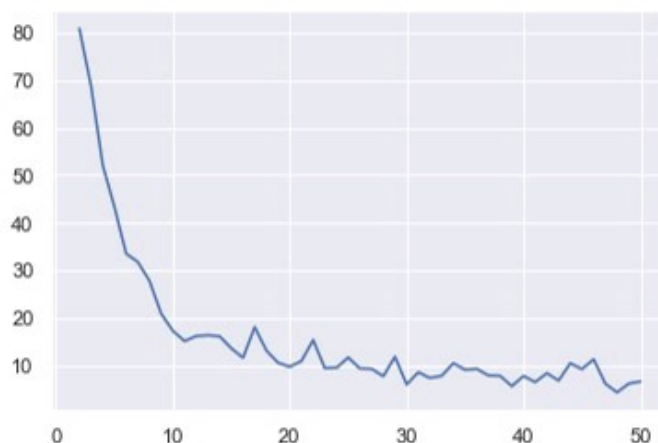


Figura 5.27: Valores obtidos através dos método Elbow (Eixo Y) para cada valor de “k” (Eixo X).

A distribuição do número de usuários que cada comunidade possui está listada na Tabela 5.9.

Tabela 5.9: Relação de comunidades identificadas pelo K-Means e a sua quantidade de usuários.

Comunidade	Número de Usuários
01	129
02	277
03	107
04	44
05	224
06	51
07	187
08	72
09	87
10	186
Média	136.4

5.4.2.2 Classificação dos usuários em comunidades utilizando DBSCAN

Para realizar a classificação dos usuários em comunidades utilizando o DBSCAN, ao contrário do K-Means, não é necessário definir um número “k” de *clusters* a priori da execução do algoritmo. Entretanto, é necessário a definição de duas outras variáveis: (a) *eps* e (b) *min_samples*.

A variável *eps*, como detalhado na Seção 2.1.2.4, refere-se à distância máxima entre dois registros para que ambos sejam considerados como parte de um mesmo *cluster*, ou comunidade. Já a variável *min_samples* é a quantidade mínima de

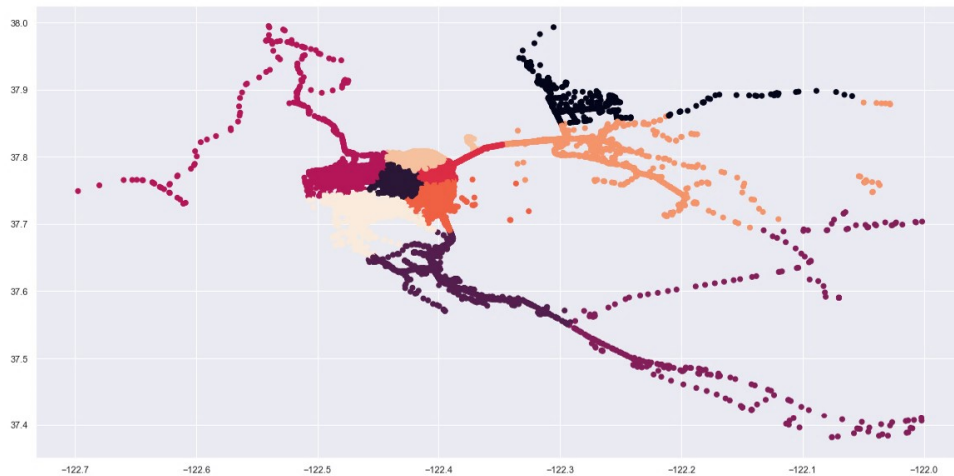


Figura 5.28: Distribuição espacial dos registros classificados em comunidades utilizando o K-Means.

registros dentro de uma mesma comunidade, de forma que esta seja validada como uma comunidade real e não apenas como ruído nos dados e, portanto, desconsiderada.

Não é possível realizar a definição da variável `min_samples` de uma forma eficiente. Sua definição depende do conhecimento prévio da fonte de dados e das características das comunidades que se pretende formar. Sendo assim, de acordo com a análise prévia dos dados do *dataset* utilizado, como quantidade de registros e sua distribuição, definiu-se para este experimento o valor desta variável como `min_samples = 1000`.

Já para a variável `eps` existem métodos que auxiliam em sua escolha, de maneira a facilitar sua configuração. Estes métodos consistem em elencar, através do cálculo de distância entre os elementos, um ranqueamento dos elementos de forma a identificar o valor ótimos de `eps`. Este processo é bem detalhado e explicado em [46] e [47].

Desta forma, utilizou-se o algoritmo KNN para verificar o ranqueamento das distâncias entre os registros de forma a obter um valor ótimo aproximado para `eps`. Os valores obtidos após a execução do KNN são apresentados na Figura 5.29.

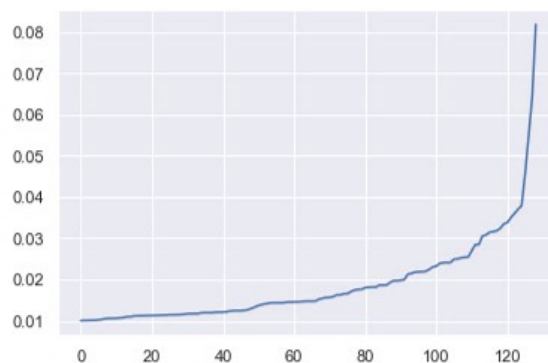


Figura 5.29: Valores obtidos após a execução do KNN para o ranqueamento entre as distâncias entre os registros.

A partir da análise dos valores resultantes do processamento do KNN, foi definido `eps` com o valor `eps = 0.035`. Com ambos valores definidos, `eps` e `min_samples`, é executado o DBSCAN para a classificação de usuários em comunidades. Como resultado da classificação dos usuários em comunidades foram identificadas 9 comunidades, sendo a distribuição dos usuários conforme a Figura 5.30.

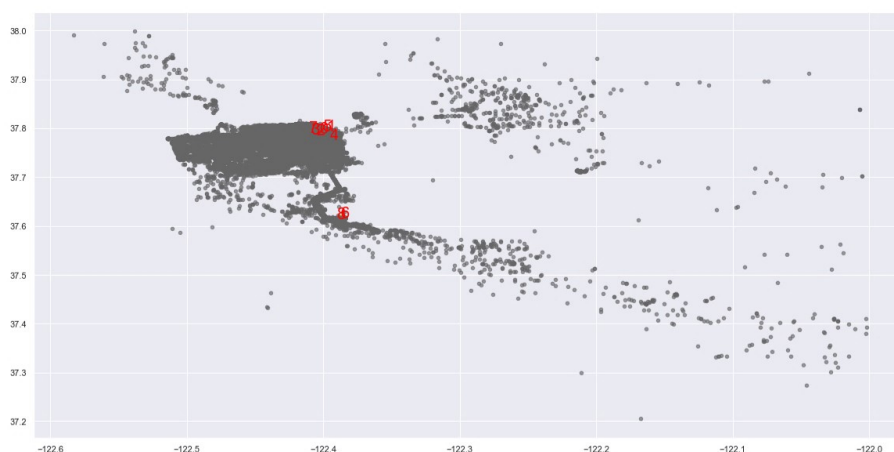


Figura 5.30: Distribuição espacial dos registros classificados em comunidades utilizando o DBSCAN.

É possível notar na Figura 5.30 que as comunidades identificadas estão concentradas muito próximas umas das outras. Isto se deve à característica do algoritmo de

clusterização DBSCAN na identificação das comunidades. A distribuição do número de usuários que cada comunidade identificada possui está lista na Tabela 5.10.

Tabela 5.10: Relação de comunidades identificadas utilizando o DBSCAN e suas respectivas quantidades de usuários.

Comunidade	Número de Usuários
01	412
02	406
03	365
04	413
05	347
06	352
07	393
08	283
09	289
Média	373.3

5.4.3 Definição de eventos

Neste experimento, assim como nos demais experimentos demonstrados, os eventos definidos são escolhidos de forma que cada local de evento possua uma semântica para usuários que com ele interagem.

Os locais selecionados para a ocorrência dos eventos são listados abaixo:

1. **Palácio de Finas Artes** (37.80469777209957, -122.44794354867301)
2. **Fonte Vaillancourt** (37.79556234680666, -122.3950997486141)
3. **Rua famosa pelas casas vitorianas** (37.7780996108072, -122.4316432257146)
4. **Universidade da Califórnia em São Francisco, Parnassus Campus** (37.762784130191676, -122.45781709647646)
5. **Hospital Geral de São Francisco** (37.75589805232178, -122.40501108184367)
6. **Museu Memorial de M. H. de Young** (37.771292471229685, -122.46857693746166)
7. **Museu de Arte Moderna de São Francisco** (37.78674207404388, -122.40146202250489)
8. **Coit Tower** (37.803478931601084, -122.40550648490225)
9. **Potrero Business Center** (37.750893252241674, -122.39420263425968)
10. **Pirâmide Transamerica** (37.79499083899438, -122.40422109750119)

Para cada um destes onze locais de eventos apresentados acima, sete cenários de evento serão analisados. Estes cenários são listados a seguir:

1. Eventos ocorrendo às Segundas-feiras;
2. Eventos ocorrendo às Terças-feiras;
3. Eventos ocorrendo às Quartas-feiras;
4. Eventos ocorrendo às Quintas-feiras;
5. Eventos ocorrendo às Sextas-feiras;
6. Eventos ocorrendo aos Sábados;
7. Eventos ocorrendo aos Domingos;

Os locais de eventos que foram selecionados e listados acima, são apresentados na Figura 5.31

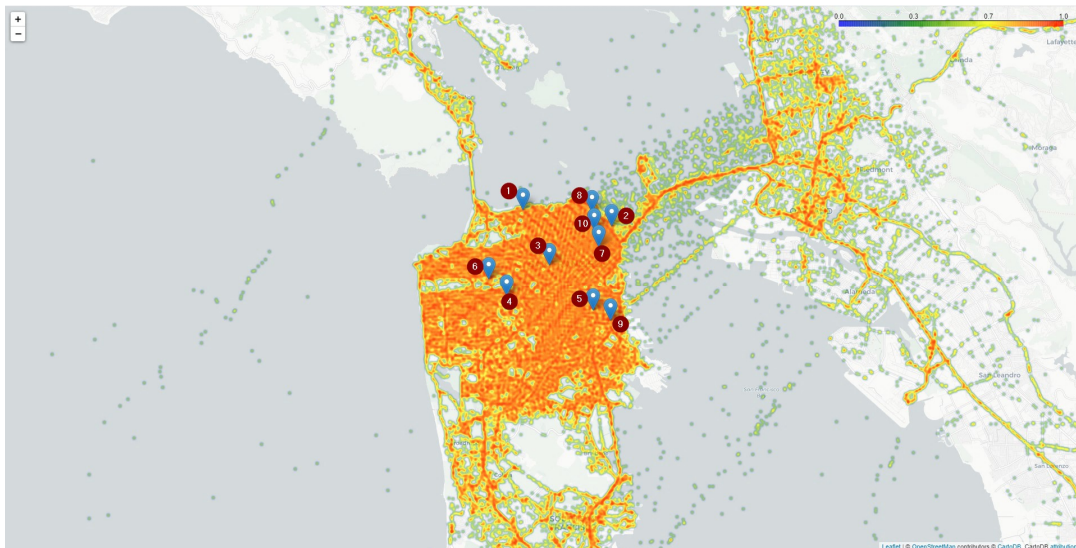


Figura 5.31: Locais de eventos selecionados apresentados sobre o mapa da cidade de São Francisco.

Além dos locais dos eventos, definiu-se a área de efeito de todos eventos como sendo a área de uma circunferência de raios 100 metros. Já para o fator de influência

(IF), IF será definido com os valores apresentados na Figura 5.32, que foi criada a partir de uma discretização de todos os dados disponíveis no *dataset* do CABSPOTTING que foram apresentados na figura de distribuição temporal apresentados na Figura 5.9.

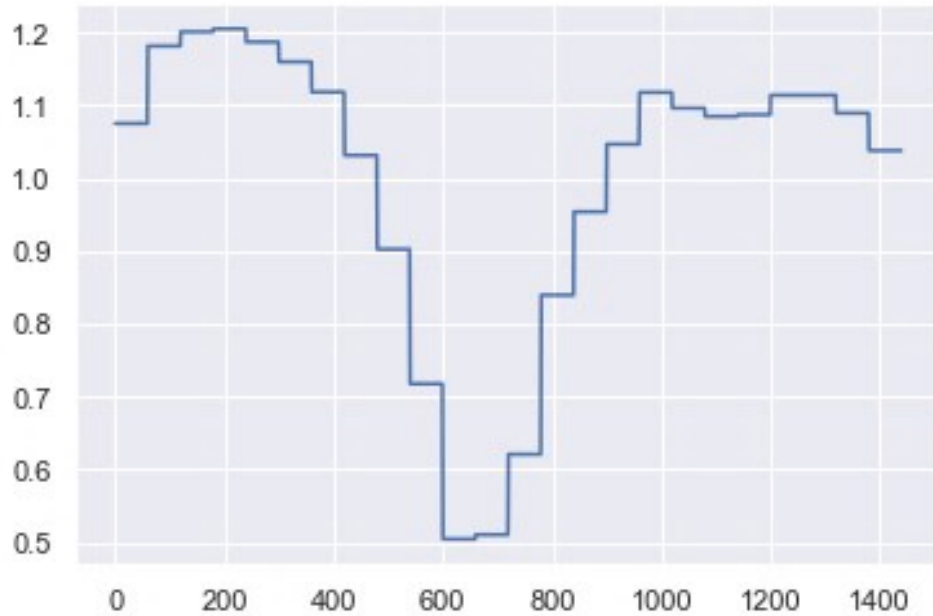


Figura 5.32: Fator de Influência IF (Eixo Y) em função do tempo (Eixo X).

O Fator de Influência, como explicado na seção 3.3.4 deve ser gerado a partir de uma análise prévia dos eventos. Neste experimento, utilizamos a discretização dos dados da distribuição temporal dos usuários ao longo de 24 horas de forma a sugerir que os períodos nos quais possuem maior movimentação de usuários coincidem com uma maior influência de um evento sobre os usuários em decorrência desta maior movimentação de usuários.

5.4.4 Apresentação dos Resultados

Para cada um dos dez locais de eventos, calculou-se a influência para cada semana dentro de todo período do *dataset*. Com a combinação dos dez locais de eventos com todos os sete cenários listados anteriormente, foram realizadas as análises de ao todo 70 possibilidades de eventos.

Cada nome de evento é composto por um identificador de evento sequencial, seguido pela letra "E" e o identificador do local de evento, seguido por três letras para a identificação do dia da semana. A título de exemplo, um evento com o nome

“01 - E01 - MON” representa um evento de identificador “01” que ocorre no local de evento 01 (Palácio de Finas Artes) no dia da semana “Segunda-feira”.

Para a análise da influência de eventos sobre comunidades, neste experimento foram utilizadas comunidades geradas de duas formas: (a) utilizando K-Means e (b) utilizando DBSCAN. As comunidades geradas utilizando cada um desses dois métodos serão analisadas de duas formas distintas: (a) a primeira realizando a comparação de eventos que ocorrem no mesmo local, de forma a analisar a variação da influência ao longo da semana, facilitando a comparação dia a dia; e (b) a segunda realizando a comparação de eventos que ocorrem em diferentes locais, porém em um mesmo dia da semana e, assim, facilitando a comparação da influência de diferentes locais dado um mesmo período de ocorrência deste evento.

5.4.4.1 Apresentação dos Resultados - K-Means

As comparações de um mesmo evento em diferentes dias da semana são utilizadas com o objetivo de identificar semelhanças e diferenças no comportamento de eventos através da análise da influência gerada sobre comunidades. Na Figura 5.33 é apresentado um gráfico para cada um dos onze locais de evento. Cada gráfico representa a influência daquele evento para cada uma das dez comunidades estudadas, em cada dia da semana. Já na Figura 5.34 são apresentados gráficos para cada dia da semana, mostrando todos os dez eventos. Cada gráfico apresenta a influência obtida de cada evento para cada uma das dez comunidades.

Como pode ser visto, na Figura 5.33, alguns eventos influenciam com maior intensidade algumas comunidades quando comparadas com outras. Os locais de evento “E7” e “E10” são os locais que afetam comunidades com maior intensidade. Os locais de evento “E2”, “E7” e “E10” possuem uma tendência de influenciar as comunidades com maior intensidade durante dias úteis se comparado com finais de semana. Por outro lado, os locais de evento “E3” e “E6” tendem a influenciar as comunidades com maior intensidade durante finais de semana.

Como pode ser visto, na Figura 5.34, eventos afetam comunidades de formas diferentes mesmo durante os mesmos dias da semana. Esta característica se deve provavelmente ao fato da existência de diferentes características de comportamento entre essas comunidades formadas a partir do K-Means ao longo da semana.

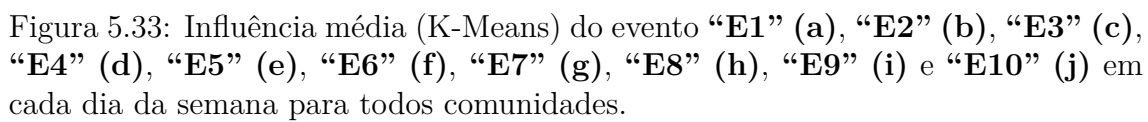
5.4.4.2 Apresentação dos Resultados - DBSCAN

Assim como no K-Means, foram realizadas comparações de um mesmo evento em diferentes dias da semana que são utilizadas com o objetivo de identificar semelhanças e diferenças no comportamento de eventos através da análise da influência gerada sobre comunidades.

Na Figura 5.35, é apresentado um gráfico para cada um dos dez locais de eventos. Cada gráfico representa a influência daquele evento para cada uma das nove comunidades estudadas, uma para cada dia da semana. Já na Figura 5.36, são apresentados gráficos para cada dia da semana mostrando todos os dez eventos. Cada gráfico apresenta a influência obtida de cada evento para cada uma das nove comunidades.

Como pode ser visto nas Figuras 5.35 e 5.36, eventos que ocorrem no mesmo local influenciam comunidades de forma diferentes quando comparado em dias diferentes. Além disso, é possível também observar que o mesmo acontece quando se compara eventos que ocorrem em locais diferentes no mesmo dia. Estas características estão relacionadas com a mobilidade que cada comunidade possui.

Porém, ao contrário do que pode ser verificado na análise de influência realizada para as comunidades geradas utilizando o K-Means, as comunidades geradas pelo DBSCAN não possuem diferenças significativas entre si. Desta forma, fica claro que a classificação de usuários em comunidades, para o critério definido e utilizando o DBSCAN, não contribui para uma boa “distinção” entre as comunidades formadas. Pode-se dizer então que, para este caso, uma vez que não há diferenças significativas entre os níveis de influência dessas comunidades, a forma de obtenção das mesmas através do DBSCAN não mapeou semânticas que gerem diferenças significativas de comportamento entre essas comunidades. Desse modo, o uso do DBSCAN inviabilizou a adoção de estratégias que possam ser direcionadas para comunidades específicas neste cenário.



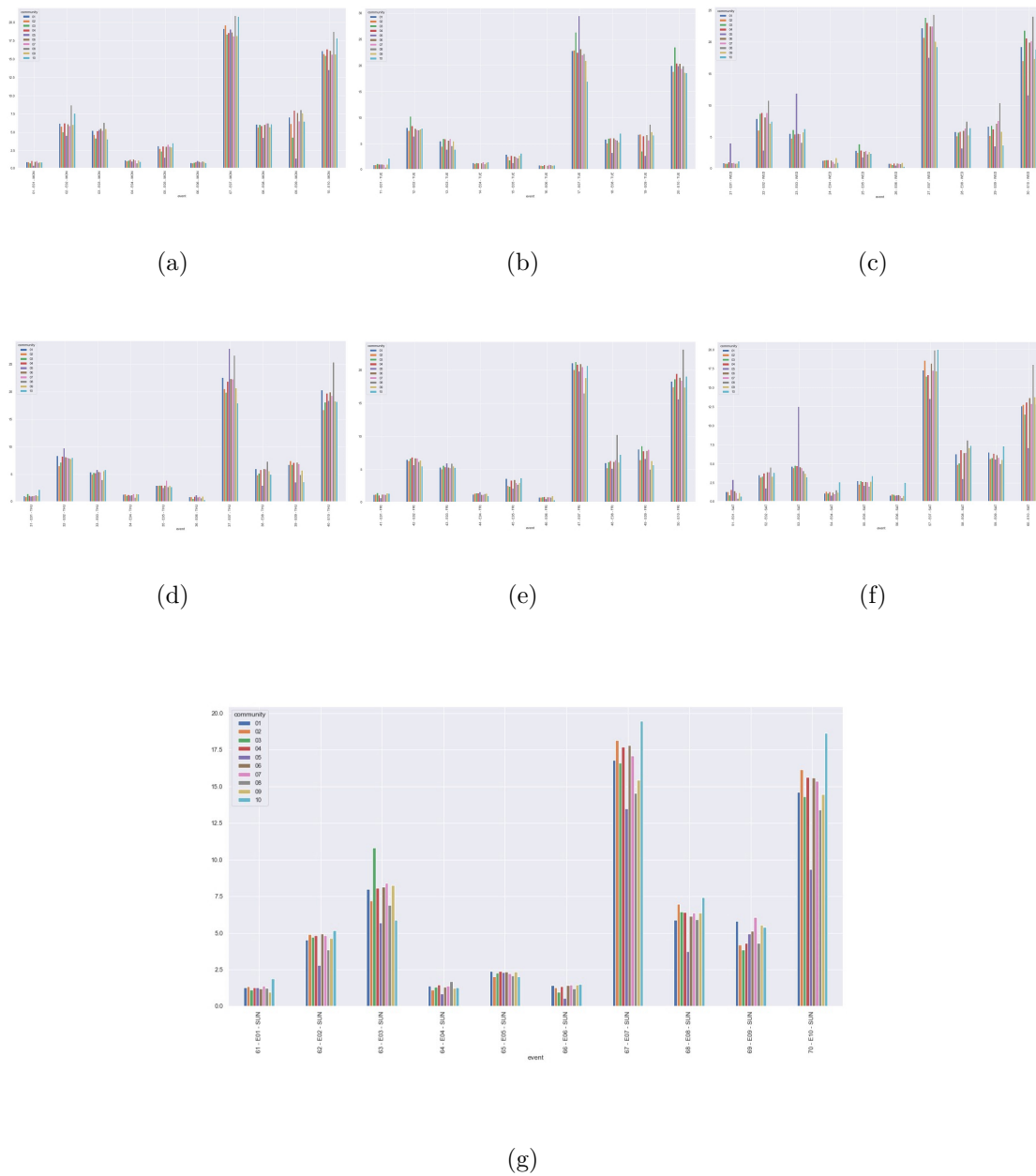


Figura 5.34: Influência média (K-Means) de todos dez eventos na **Segunda** (a), na **Terça** (b), na **Quarta** (c), na **Quinta** (d), na **Sexta** (e), no **Sábado** (f) e no **Domingo** (g).

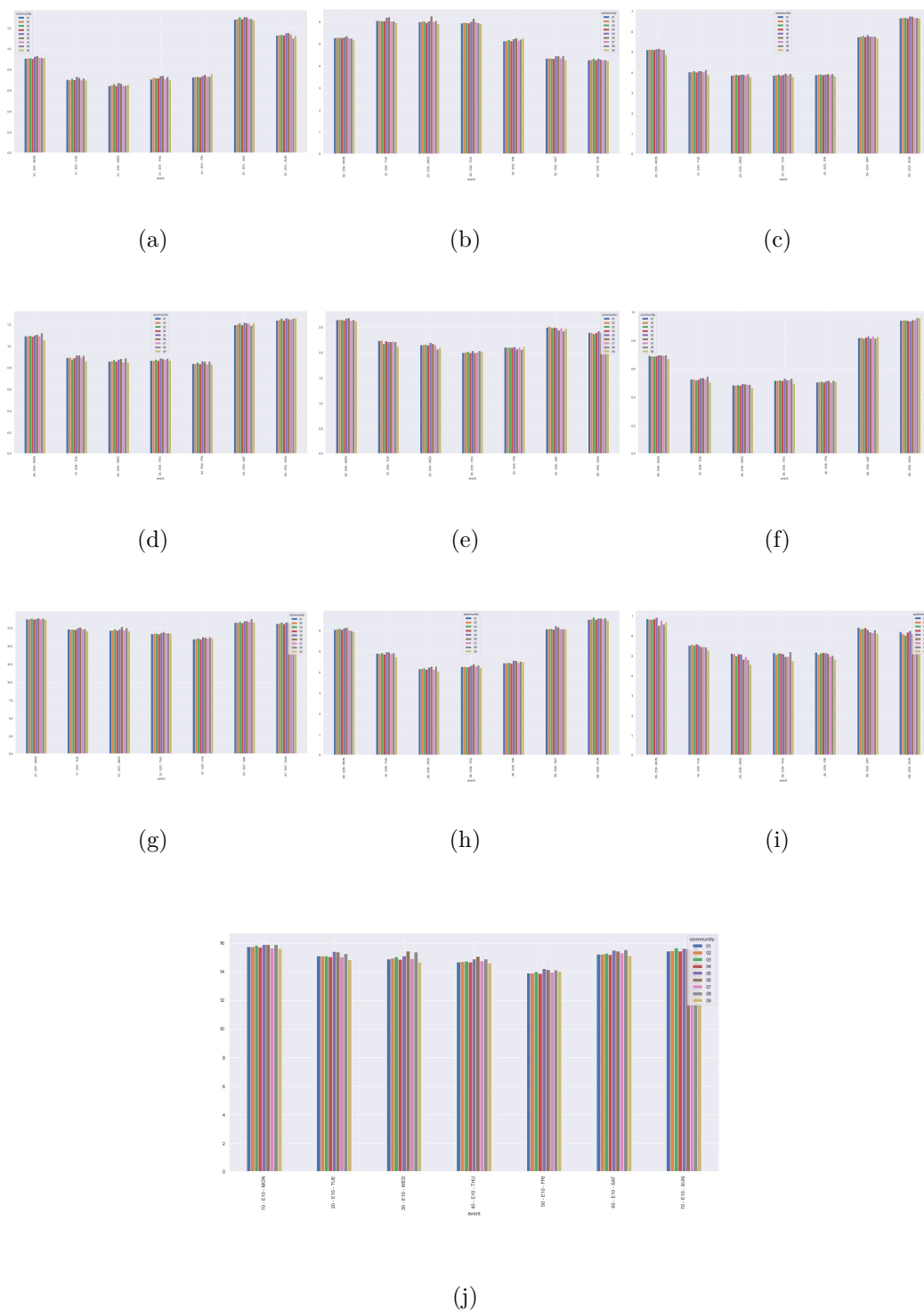


Figura 5.35: Influência média (DBSCAN) do evento “E1” (a), “E2” (b), “E3” (c), “E4” (d), “E5” (e), “E6” (f), “E7” (g), “E8” (h), “E9” (i) e “E10” (j) em cada dia da semana para todos comunidades.

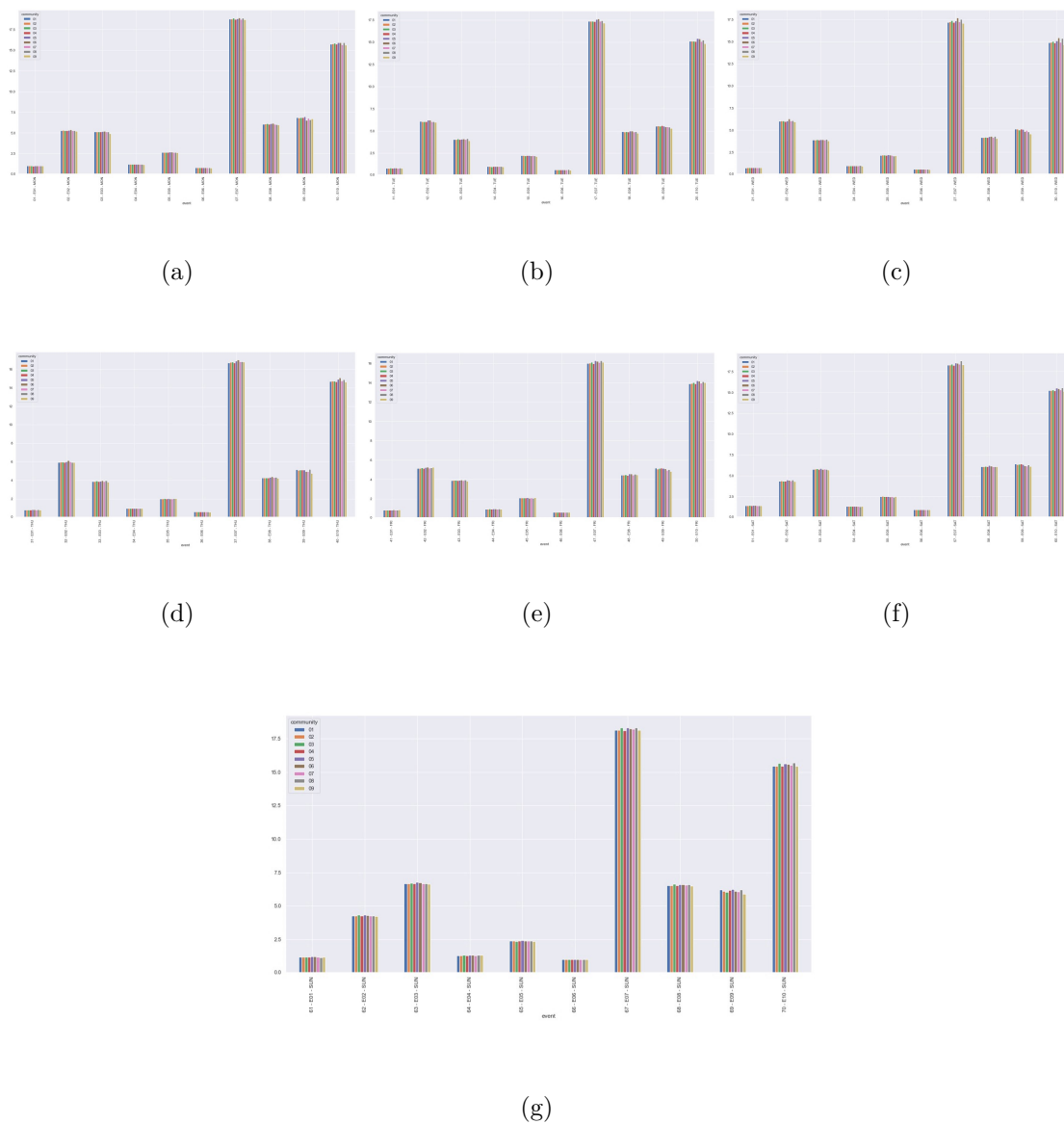


Figura 5.36: Influência média (DBSCAN) de todos dez eventos na **Segunda** (a), na **Terça** (b), na **Quarta** (c), na **Quinta** (d), na **Sexta** (e), no **Sábado** (f) e no **Domingo** (g).

6. Conclusão e Trabalhos Futuros

Neste trabalho foi endereçado o problema de análise da influência de eventos sobre comunidades utilizando dados de mobilidade urbana e como esses eventos interagem com comunidades e seus usuários. Baseado na análise de trabalhos relacionados, verificou-se que a análise de eventos é realizada de forma muito específica com pouco grau de generalização. Além disso, a análise desses eventos, quando realizada, considera apenas a influência desses eventos sobre usuários como indivíduos e não como parte de um grupo que compartilham entre si características e comportamentos semelhantes.

Desta forma, foi aqui proposta uma metodologia que, além de generalizar a questão da influência de eventos sem comprometer a análise das particularidades de cada um desses eventos, realiza uma análise abrangente envolvendo comunidades de usuários e os efeitos desses eventos sobre elas.

A metodologia proposta é composta por um fluxo de ações que envolve o tratamento dos dados de mobilidade, necessário para que estes possam ser processados, e descreve métodos para a classificação de usuários e consequente formação de comunidades que possuam semânticas atreladas a elas. Além disso, define também as principais características que um evento deve possuir, como a área onde o evento exerce influência, tempo e período de duração, localização do evento etc. Além disso, são descritas as equações para o cálculo efetivo das medidas de influência desses eventos sobre comunidades ou sobre os indivíduos dessas comunidades, aqui tratados pelo termo “usuários”.

Através dos três experimentos desenvolvidos, notou-se que a partir da obtenção de comunidades com semânticas bem definidas é possível propor soluções voltadas para comunidades específicas, de forma que se possa otimizar o planejamento de diversas ações de interesse público, como de saúde ou sanitárias (por exemplo, uma

campanha de vacinação), ou ainda de publicidade (por exemplo, uma campanha de *marketing*). Foi possível entender também a importância da utilização de comunidades criadas com semânticas claras e utilizando métodos apropriados, como demonstrado no experimento 3, onde foi demonstrado que comunidades criadas com semânticas parecidas não geram valor para a aplicação da metodologia proposta.

Apesar de terem sido apresentados experimentos e definições de comunidades e eventos que utilizam semântica, entende-se que a proposta apresentada pode ser estendida para aplicações nas quais não haja a necessidade de uma semântica atrelada a eventos e comunidades.

Com todo estudo realizado neste trabalho, obteve-se conhecimento de como eventos e comunidades interagem entre si e se relacionam. Foi possível também validar a equação proposta para o cálculo da medida de influência de eventos sobre comunidades, aplicando-a para três experimentos distintos.

Como trabalhos futuros, busca-se a aplicação de todo a metodologia desenvolvida até então em uma fonte de dados de mobilidade densa, gerada através da simulação de cenários realistas e representativos de uma sociedade real. Além disso, pretende-se utilizar novas características para a classificação de usuários em comunidades, principalmente aquelas orientadas a modais de transporte distintos e a características geográficas distintas, como locais de trabalho, de estudo, de lazer etc.

Além disso, pretende-se analisar o comportamento da influência e das comunidades antes e após as ocorrências de eventos, de forma a analisar se existe uma retroalimentação do modelo, verificando assim a existência de uma influência duradora de eventos sobre comunidades e modificando a mobilidade destas comunidades.

Pretende-se também analisar a relação de influência de eventos de maneira espacial, ou seja, verificar se existem relacionamentos entre locais de eventos utilizando a influência como fator de comparação.

Por fim, pretende-se também desenvolver métodos de criação de comunidades utilizando como critério a influência de eventos sobre os usuários, de forma a identificar se existe relação entre a influência de eventos e os atributos dos usuários. Na hipótese positiva, pretende-se analisar a relação entre esses atributos e a influência com que eventos afetam os respectivos usuários, realizando uma análise estatística para comparar diferentes formas de agrupamento.

Referências Bibliográficas

- [1] A. Saxena, M. Prasad, A. Gupta, N. Bharill, O. P. Patel, A. Tiwari, M. J. Er, W. Ding, and C.-T. Lin, “A review of clustering techniques and developments,” *Neurocomputing*, vol. 267, pp. 664–681, 2017.
- [2] K. D. Bess, A. T. Fisher, C. C. Sonn, and B. J. Bishop, *Psychological Sense of Community: Theory, Research, and Application*, pp. 3–22. Springer, 2002.
- [3] Y. Zheng, H. Fu, X. Xie, W.-Y. Ma, and Q. Li, *Geolife GPS trajectory dataset - User Guide*, geolife gps trajectories 1.1 ed., July 2011. Geolife GPS trajectories 1.1.
- [4] M. Piorkowski, N. Sarafijanovoc-Djukic, and M. Grossglauser, “A Parsimonious Model of Mobile Partitioned Networks with Clustering,” in *The First International Conference on COMmunication Systems and NETworkS (COMSNETS)*, January 2009.
- [5] A. K. Jain, M. N. Murty, and P. J. Flynn, “Data clustering: A review,” *ACM Comput. Surv.*, vol. 31, p. 264–323, Sept. 1999.
- [6] A. Mohebi, S. Aghabozorgi, T. Ying Wah, T. Herawan, and R. Yahyapour, “Iterative big data clustering algorithms: a review,” *Software: Practice and Experience*, vol. 46, no. 1, pp. 107–129, 2016.
- [7] B. Xue, M. Zhang, W. N. Browne, and X. Yao, “A survey on evolutionary computation approaches to feature selection,” *IEEE Transactions on Evolutionary Computation*, vol. 20, no. 4, pp. 606–626, 2016.
- [8] C. Cooper, D. Franklin, M. Ros, F. Safaei, and M. Abolhasan, “A comparative survey of vanet clustering techniques,” *IEEE Communications Surveys Tutorials*, vol. 19, no. 1, pp. 657–681, 2017.

- [9] S. Guha, R. Rastogi, and K. Shim, “Rock: A robust clustering algorithm for categorical attributes,” *Information systems*, vol. 25, no. 5, pp. 345–366, 2000.
- [10] C. Yabing and C. Gao, “Birch: An efficient clustering method for very large databases,” 1999.
- [11] S. Guha, R. Rastogi, and K. Shim, “Cure: An efficient clustering algorithm for large databases,” *ACM Sigmod record*, vol. 27, no. 2, pp. 73–84, 1998.
- [12] S. Nanjundan, S. Sankaran, C. Arjun, and G. P. Anand, “Identifying the number of clusters for k-means: A hypersphere density based approach,” *arXiv preprint arXiv:1912.00643*, 2019.
- [13] A. Satre-Meloy, M. Diakonova, and P. Grünewald, “Cluster analysis and prediction of residential peak demand profiles using occupant activity data,” *Applied Energy*, vol. 260, p. 114246, 2020.
- [14] D. L. Ferreira, *Characterization of Human Social Mobility Patterns Applied to Mobility Modelling and Opportunistic Networks*. PhD thesis, UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO, 2019.
- [15] S. Aggarwal, N. Agarwal, and M. Jain, “Performance analysis of uncertain k-means clustering algorithm using different distance metrics,” in *Computational Intelligence: Theories, Applications and Future Directions - Vol I*, pp. 237–245, Springer, 2019.
- [16] G. Ogbuabor and F. Ugwoke, “Clustering algorithm for a healthcare dataset using silhouette score value,” *International Journal of Computer Science & Information Technology*, vol. 10, no. 2, pp. 27–37, 2018.
- [17] E. SCHUBERT, J. SANDER, M. ESTER, H.-P. KRIEGEL, and X. XIAOWEI, “Dbscan revisited, revisited: Why and how you should (still) use dbscan.,” *ACM Transactions on Database Systems*, vol. 42, no. 3, pp. 1 – 21, 2017.
- [18] P. D. McNicholas, “Model-based clustering,” *Journal of Classification*, vol. 33, pp. 331–373, Oct 2016.
- [19] W. Zhang, R. Li, P. Shang, and H. Liu, “Impact analysis of rainfall on traffic flow characteristics in beijing,” *International Journal of Intelligent Transportation Systems Research*, vol. 17, pp. 150–160, May 2019.

- [20] M. Pregnotato, A. Ford, S. M. Wilkinson, and R. J. Dawson, “The impact of flooding on road transport: A depth-disruption function,” *Transportation Research Part D: Transport and Environment*, vol. 55, pp. 67 – 81, 2017.
- [21] T. Maze, M. Agarwai, and G. Burchett, “Whether weather matters to traffic demand, traffic safety, and traffic operations and flow,” *Transportation Research Record*, vol. 1948, pp. 170–176, 01 2006.
- [22] L. Marín-León and M. M. Vizzotto, “Comportamentos no trânsito: um estudo epidemiológico com estudantes universitários,” *Cadernos de Saúde Pública*, vol. 19, pp. 515–523, 2003.
- [23] K. C. Roy, M. Cebrian, and S. Hasan, “Quantifying human mobility resilience to extreme events using geo-located social media data,” *EPJ Data Science*, vol. 8, p. 18, May 2019.
- [24] A. T. Hojati, L. Ferreira, S. Washington, P. Charles, and A. Shobeirinejad, “Reprint of: Modelling the impact of traffic incidents on travel time reliability,” *Transportation Research Part C: Emerging Technologies*, vol. 70, pp. 86 – 97, 2016.
- [25] D. L. Ferreira, C. de Souza, K. Obraczka, and C. A. V. Campos, “Identifying user communities using deep learning and its application to opportunistic networking,” in *16th IEEE International Conference on Mobile Ad Hoc and Sensor Systems, MASS 2019, Monterey, CA, USA, November 4-7, 2019*, pp. 344–352, IEEE, 2019.
- [26] H. Feng, J. Tian, H. J. Wang, and M. Li, “Personalized recommendations based on time-weighted overlapping community detection,” *Information & Management*, vol. 52, no. 7, pp. 789 – 800, 2015. Novel applications of social media analytics.
- [27] D. Stoltenberg, D. Maier, and A. Waldherr, “Community detection in civil society online networks: Theoretical guide and empirical assessment,” *Social Networks*, vol. 59, pp. 120 – 133, 2019.
- [28] X. Li, C. Sun, and M. A. Zia, “Social influence based community detection in event-based social networks,” *Information Processing & Management*, vol. 57, no. 6, p. 102353, 2020.

- [29] M. A. Hoque, X. Hong, and B. Dixon, “Analysis of mobility patterns for urban taxi cabs,” in *2012 International Conference on Computing, Networking and Communications (ICNC)*, pp. 756–760, 2012.
- [30] F. Rehman, O. Khalid, and S. A. Madani, “A comparative study of location-based recommendation systems.,” *Knowledge Eng. Review*, vol. 32, p. e7, 2017.
- [31] M. Narayanan and A. K. Cherukuri, “A study and analysis of recommendation systems for location-based social network (lbsn) with big data,” *IIMB Management Review*, vol. 28, no. 1, pp. 25 – 30, 2016.
- [32] E. Toch, B. Lerner, E. Ben-Zion, and I. Ben-Gal, “Analyzing large-scale human mobility data: a survey of machine learning methods and applications.,” *Knowledge & Information Systems*, vol. 58, no. 3, pp. 501 – 523, 2019.
- [33] M. C. da Mata Martins, A. N. Rodrigues da Silva, and N. Pinto, “An indicator-based methodology for assessing resilience in urban mobility,” *Transportation Research Part D: Transport and Environment*, vol. 77, pp. 352 – 363, 2019.
- [34] S. Wang, L. Li, W. Ma, and X. Chen, “Trajectory analysis for on-demand services: A survey focusing on spatial-temporal demand and supply patterns,” *Transportation Research Part C: Emerging Technologies*, vol. 108, pp. 74 – 99, 2019.
- [35] B. B. Majumdar and S. Mitra, “Analysis of bicycle route-related improvement strategies for two indian cities using a stated preference survey,” *Transport Policy*, vol. 63, pp. 176 – 188, 2018.
- [36] M. Saphioğlu and M. Aydın, “Choosing safe and suitable bicycle routes to integrate cycling and public transport systems,” *Journal of Transport & Health*, vol. 10, pp. 236 – 252, 2018.
- [37] A. A. Campbell, C. R. Cherry, M. S. Ryerson, and X. Yang, “Factors influencing the choice of shared bicycles and shared electric bikes in beijing,” *Transportation research part C: emerging technologies*, vol. 67, pp. 399–414, 2016.
- [38] D. Sathiaraj, T. on Punksam, F. Wang, and D. P. Seedah, “Data-driven analysis on the effects of extreme weather elements on traffic volume in atlanta, ga, usa,” *Computers, Environment and Urban Systems*, vol. 72, pp. 212 – 220, 2018.
- [39] S. Kwoczek, *Enhanced mobility awareness: a data-driven approach to analyze traffic under planned special event scenarios*. PhD thesis, Hannover: Institutionelles Repositorium der Leibniz Universität Hannover, 2018.

- [40] Â. Corrêa and H. Sferra, “Conceitos e aplicações de data mining,” *Revista de ciência & tecnologia*, vol. 11, pp. 19–34, 2003.
- [41] S. Zhang, C. Zhang, and Q. Yang, “Data preparation for data mining,” *Applied Artificial Intelligence*, vol. 17, no. 5-6, pp. 375–381, 2003.
- [42] D. L. Ferreira, B. A. A. Nunes, C. A. V. Campos, and K. Obraczka, “A deep learning approach for identifying user communities based on geographical preferences and its applications to urban and environmental planning,” *ACM Trans. Spatial Algorithms Syst.*, vol. 6, no. 3, pp. 17:1–17:24, 2020.
- [43] K. M. Vespucci, *Operação horário de pico*. CET, 2005.
- [44] M. Souza, S. Lucena, and C. Campos, “Uma proposta para obter o grau de influência de eventos sobre comunidades baseadas em critérios de geolocalização,” in *Anais do IV Workshop de Computação Urbana*, (Porto Alegre, RS, Brasil), pp. 111–124, SBC, 2020.
- [45] L. Qiao, Y. Shi, and S. Chen, “An empirical study on the temporal structural characteristics of vanets on a taxi gps dataset,” *IEEE Access*, vol. 5, pp. 722–731, 2017.
- [46] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, “A density-based algorithm for discovering clusters in large spatial databases with noise,” in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD’96, p. 226–231, AAAI Press, 1996.
- [47] N. Rahmah and I. Sitanggang, “Determination of optimal epsilon (eps) value on dbscan algorithm to clustering data on peatland hotspots in sumatra,” *IOP Conference Series: Earth and Environmental Science*, vol. 31, p. 012012, 01 2016.