Uma proposta para obter o grau de influência de eventos sobre comunidades baseadas em critérios de geolocalização

Marcos A. de P. Souza[1], **Sidney C. de Lucena**[1], **Carlos A. V. Campos**[1]

[1]Universidade Federal do Estado do Rio de Janeiro (UNIRIO)
Av. Pasteur, 458 – Rio de Janeiro – RJ – Brasil

Abstract. *Human society is divided into communities formed by individuals who share similarities in behavior, interests, housing, etc. Events such as advertising actions or traffic accidents, among others, occur daily in urban centers and affect these communities differently. This article proposes an approach to assess the influence of the location of these events on different communities, thus contributing to the planning of actions aimed at these communities. To this end, a case study was carried out using GEOLIFE, in which users were divided into communities based on the history of geolocation, for which the degree of influence of events occurred in eleven places of great circulation was calculated. The application of these results was demonstrated by maximizing the effectiveness of a vaccination campaign with a limited budget in the city of Beijing.*

Resumo. *A sociedade humana divide-se em comunidades formadas por indivíduos que compartilham similaridades de comportamento, interesses, moradia etc. Eventos como ações publicitárias ou acidentes de trânsitos, dentre outros, ocorrem diariamente nos centros urbanos e afetam de forma diferente essas comunidades. Este artigo propõe uma abordagem para avaliar a influência do local de ocorrência desses eventos sobre diferentes comunidades, contribuindo assim com o planejamento de ações voltadas para essas comunidades. Para isso, realizou-se um estudo de caso usando o dataset GEOLIFE, no qual os usuários foram divididos em comunidades baseadas no histórico de geolocalização, para as quais se calculou o grau de influência de eventos ocorridos em onze locais de grande circulação. A aplicação desses resultados foi demonstrada maximizando-se a eficácia de uma campanha de vacinação com orçamento limite na cidade de Pequim.*

1. Introdução

A sociedade humana é formada e organizada em inúmeras comunidades, com as mais diversas características e interesses. Segundo [Bess et al. 2002], existem dois tipos de comunidades: as baseadas em localidade e as baseadas em relacionamento, formadas a partir de interesses em comum, problemas em comum ou características em comum.

Ainda segundo [Bess et al. 2002], cada pessoa pertence a mais de uma comunidade e as comunidades às quais pertencemos podem e provavelmente irão variar ao longo do tempo e das circunstâncias. Podemos então afirmar que comunidades podem ser formadas a partir de diversas características, como torcer para um time de futebol ou por alinhamento político-ideológico.

Em um mundo globalizado como o atual, essas comunidades interagem entre si e compartilham cotidianamente locais de trabalho, sistemas de transporte, locais de lazer,

regiões de domicílios, experiências ou informações. Além disso, por possuírem características e padrões de mobilidade diferentes, essas comunidades estão sujeitas a graus de alcance diferentes mediante a ocorrência de eventos em determinada região. Em outras palavras, o impacto de um evento pode ter diferentes alcances para cada comunidade impactada. Além disso, eventos e comunidades precisam ter alguma relação entre si para que um influencie o outro. Por exemplo, em uma comunidade rural, eventos exclusivamente urbanos podem ter pouca ou nenhuma relevância.

Este artigo tem como objetivo avaliar o grau de influência do local de um determinado evento sobre diferentes comunidades organizadas por critério de moradia. Para tal, é proposta uma abordagem para identificar o efeito de possíveis eventos que afetem regiões predominantemente urbanas onde existe um alto nível de movimentação de pessoas e veículos. São localidades onde podem ocorrer, diariamente, eventos como ações publicitárias em painéis digitais de prédios ou de pontos de ônibus, protestos populares, eventos culturais em praças, jogos em estádios, manutenções viárias, acidentes entre veículos ou problemas decorrentes de fenômenos climáticos, dentre outros.

Ressalta-se que uma das possíveis aplicações deste trabalho, ao avaliar o efeito de eventos em comunidades, é o fornecimento de informações relevantes para diversas áreas, como *marketing*, planejamento urbano estratégico ou análise de grupos de risco de doenças. Nas ações de marketing, por exemplo, o conhecimento adquirido pode auxiliar na utilização de painéis dinâmicos em horário desejado, considerando a disposição de seu público-alvo. Como consequência, obtém-se maior eficiência nessas ações e minimiza-se a poluição visual nas cidades. Já na área de saúde pública, a análise comportamental e de movimentação de um ou mais grupos de risco de uma determinada doença permite identificar potenciais locais para campanhas de prevenção. [Feng et al. 2015] corrobora com essa hipótese ao afirmar que, conhecendo-se as preferências dos indivíduos de uma dada comunidade, melhor será a efetividade das recomendações direcionadas a esses indivíduos.

Conforme nosso objetivo, foi realizado um estudo de caso utilizando o *dataset* GEOLIFE e classificando os usuários nele registrados em comunidades. Formadas essas comunidades, foram selecionados onze locais de evento e foi estimada a influência que possíveis eventos nesses locais, em diferentes dias da semana, geraria para cada uma das comunidades formadas. Os resultados obtidos foram então utilizados para a formulação e a resolução de um problema de otimização hipotético de uma campanha de vacinação com restrições orçamentárias.

Como contribuição deste artigo, pode-se então citar:

1. Proposta de uma abordagem para identificar a influência de eventos sobre comunidades. Essa proposta viabiliza iniciativas nas áreas de sistemas de recomendação utilizando dados de mobilidade e na área de personalização baseada em preferências dos usuários.
2. Aplicação da abordagem proposta em um *dataset* real de mobilidade urbana da região metropolitana de Beijing em que foram investigados onze locais de interesse e extraídas oito comunidades de usuários.
3. Modelagem de um problema de otimização para maximizar a eficácia do uso de recursos em um campanha de vacinação com base na utilização da abordagem proposta para a identificação de comunidades no *dataset* de mobilidade utilizado.

Superando o momento introdutório, segue a organização das próximas seções deste trabalho. Na Seção 2 são descritos os trabalhos relacionados a este artigo. Na Seção 3 é descrito e detalhado o problema e a metodologia utilizada. Na Seção 4 é descrito o tratamento dos dados, a realização do experimento e são apresentados os resultados para um estudo de caso. Por fim, na Seção 5 são apresentadas as conclusões e discutidos trabalhos futuros.

2. Trabalhos relacionados

Nesta seção, são descritas algumas soluções criadas para questões relacionadas a mobilidade urbana onde foram utilizadas alguma forma de agrupamento de usuários em comunidades, além de trabalhos que tratam do impacto de eventos na mobilidade urbana.

Em [Ferreira 2019] é apresentado um novo método de classificação de usuários em comunidades baseado na similaridade de mobilidade. Em linhas gerais, nesse método, os usuários são classificados utilizando variáveis como meio de transporte, velocidade de deslocamento e posicionamento espacial. Foi então possível agrupar esses usuários em comunidades de uma forma eficaz para otimizar o roteamento de informações com um menor *overhead* através de uma rede oportunística formada pelos próprios usuários. Entretanto, o enfoque de [Ferreira 2019] não é na análise de eventos, seus impactos ou o tamanho do efeito destes nessas comunidades. Assim, a utilização dessa abordagem de classificação de usuários faz-se interessante em trabalhos futuros, pois geraria sinergias na análise dos efeitos de eventos em comunidades baseadas em padrões de mobilidade.

[Pregnolato et al. 2017] analisa o impacto de condições climáticas na mobilidade com o enfoque em enchentes. Para isso, é definido o risco de uma enchente utilizando como variáveis a probabilidade da ocorrência do evento, o impacto do evento e cada uma das propriedades do sistema. Porém, não foi proposta uma utilização das informações obtidas para nenhum propósito direto, o que o difere do nosso trabalho. Como contribuição de [Pregnolato et al. 2017] pode-se citar como é possível correlacionar variáveis, como condições climáticas, a influência na mobilidade de uma localidade.

Se por um lado [Maze et al. 2006] desenvolve a questão do impacto de condições climáticas na mobilidade urbana, por outro lado [Feng et al. 2015] trabalha a questão de identificação de comunidades. Assim, é proposto um novo método de detecção de comunidades sobrepostas, chamado pelos autores *TOTAR Framework*. Além disso, esse trabalho aborda as variações de preferências dos usuários ao longo do tempo e em diversas situações, como em diferentes condições meteorológicas. Porém, o modelo foi validado apenas utilizando dados de usuários de filmes. Como contribuição de [Feng et al. 2015], pode-se citar a questão de identificação de comunidades sobrepostas, o que difere em parte do que foi proposto em [Ferreira 2019].

Em [Stoltenberg et al. 2019] é tratada a questão da detecção de comunidades baseada em debates público/políticos. Para isso, foram utilizadas heurísticas para a criação dos grafos de relacionamentos entre as comunidades e seus membros, sendo definidos pesos e direções para cada um destes relacionamentos ou arestas. Entretanto, não foram totalmente exploradas as possibilidades citadas e não foi explorado a pluralidade de comunidades dos usuários. [Stoltenberg et al. 2019] contribui possibilitando a utilização, em trabalhos futuros, a criação de método para correlacionar comunidades baseando-se no grau de influência causado por eventos nestas comunidades.

Apesar da imensa variedade de trabalhos na área de sistemas de recomendação, na área de análise de impacto na mobilidade urbana, na área de formação de comunidades e na área de recomendação de rotas, ao melhor do nosso conhecimento não foi possível encontrar trabalhos que utilizam múltiplas características, relacionadas com comunidades, voltadas para a questão da mobilidade urbana e padrões de mobilidades diferenciados em cada comunidade.

3. Metodologia

De forma a analisar o grau de impacto de eventos em comunidades e em seus membros, é necessário primeiramente definir eventos, comunidades e grau de influência.

- *Evento* é a ocorrência de qualquer fato ou circunstancia que se deseja monitorar. Eventos podem ser, mas não são limitados a: interação de pessoas com propagandas, acidentes de trânsito, manifestações populares, jogos esportivos, etc.
- *Comunidade* é um grupo de usuários com características em comum e o agrupamento em comunidades pode ser realizado por uma ou mais características desses usuários. E cada um desses usuários pode pertencer a uma ou mais comunidades.
- *Grau de influência*, ou *de alcance*, é a fração de pessoas em uma comunidade que é atingida ou influenciada por um dado evento, conforme a área de ação desse evento.

Após terem sido apresentadas as definições necessárias, são definidas as seguintes perguntas de pesquisa:

1. Eventos afetam algumas comunidades com maior intensidade que outras?
2. É possível avaliar datas onde o impacto desses eventos sejam maiores?
3. É possível realizar recomendações baseadas no impacto desses eventos nas comunidades?

Como a maior parte da população brasileira está concentrada em zonas urbanas (cerca de 85%, de acordo com IBGE [1]) e considerando as propriedades de disseminação e acesso a informações através da internet, sobretudo do *smartphone* (cerca de 75% dos domicílios brasileiros possuíam internet em 2017, segundo o IBGE [2]), é interessante a análise de como diferentes locais de eventos podem afetar a população de formas diferentes.

Eventos podem ter impactos positivos ou negativos nas comunidades. Neste artigo, foca-se na análise da influência do local do evento de maneira a ser possível criar um modelo que possa auxiliar a maximizar os efeitos positivos de um evento, como em campanhas solidárias, de vacinação ou de prevenção de doenças orientadas a comunidades específicas. Para tal, dividiu-se a experimentação em etapas visando a obtenção de resultados e posterior análise desses resultados, seguindo o fluxo de tratamento e análise de dados similar ao *workflow* de *data mining* descrito em [Han et al. 2011].

Os dados de mobilidade utilizados são referentes à região da grande Pequim, China, por ser um grande centro urbano, característica essa que favorece a análise

[1] https://educa.ibge.gov.br/jovens/conheca-o-brasil/populacao/18313-populacao-rural-e-urbana.html

[2] https://agenciadenoticias.ibge.gov.br/agencia-sala-de-imprensa/2013-agencia-de-noticias/releases/23445-pnad-continua-tic-2017-internet-chega-a-tres-em-cada-quatro-domicilios-do-pais

pretendida. Primeiramente, os dados foram filtrados (*Seleção*) para representar essa região, depois foi realizado o tratamento desses dados com conversão de valores (*Pré-processamento*) e então realizada a classificação dos usuários em *clusters* (*Mineração*).

Para a realização do estudo de caso, foi necessário caracterizar um evento e selecionar seus locais de ocorrência. Para tanto, um evento foi definido como uma tupla formada pelo local do evento e pela data da ocorrência desse evento. Foram então selecionados onze locais de eventos que possuíam alguma semântica atrelada a eles. Ao se referir a semântica, se aspira a escolha de locais com algum significado, como hospitais, aeroportos, estações de trem ou metrô, universidades, pontos viários de grande importância, etc, e não apenas coordenadas isoladas sem significado atrelado.

Para a realização de toda a experimentação foi utilizada a linguagem *Python* combinada a diversas bibliotecas de *data science*, tanto para visualização dos dados quanto para demais cálculos.

4. Experimentação, Resultados e Análise

4.1. Análises preliminares dos dados

Neste trabalho foi utilizado o *dataset* GEOLIFE [Zheng et al. 2011]. O *dataset* possui dados coletados de 182 usuários no período de 2007 a 2012, contendo um total de 24.876.978 registros. Os registros do *dataset* possuem as seguintes informações: data/hora UTC do registro, coordenada do registro coletado (latitude, longitude e altitude), meio de transporte utilizado e o ID do usuário. Apesar do campo para meio de transporte, apenas 5.427.117 registros possui essa informação cadastrada, cerca de 21,8 % do total. Portanto, não foi utilizada esta informação.

Como primeiro passo do experimento, foi realizada uma análise estatística para determinar como os dados do *dataset* estão distribuídos. Após ser verificado que os dados em grande parte estão localizados na região da grande Pequim e com o objetivo de reduzir qualquer viés que pudesse ser inserido pelos dados esparsos, foi realizada uma filtragem de forma a enquadrar apenas esta região, delimitando os dados entre as coordenadas de latitude 39,800000 e 40,200000 e entre as coordenadas de longitude 116,100000 e 116,800000. A Figura 1 exibe os dados resultantes dessa filtragem, através de um *heat Map*, onde as coordenadas médias obtidas foram de 39,97382 e de 116,3613, com desvio padrão de 0,05310998 e de 0,07983382, para latitude e longitude, respectivamente. Isso correspondeu a uma redução de 76 e 340 vezes no desvio-padrão da latitude e da longitude, respectivamente em relação aos dados sem filtragem. Os registros resultantes dessa filtragem totalizam 18.750.290 registros.

A distribuição temporal dos dados se dá conforme a Figura 2, onde é possível notar que essa distribuição não ocorre de forma homogênea ao longo do tempo. Por esse motivo, o 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 2(b). Além dessa distribuição, é importante também ressaltar a distribuição dos usuários ao longo da semana. Esta distribuição é apresentada na Tabela 1.

4.2. Classificação dos usuários em comunidades

A primeira etapa do experimento consistiu em gerar comunidades. Como dito na Seção 3, comunidades podem ser definidas de diversas formas e com diversos significados. Neste

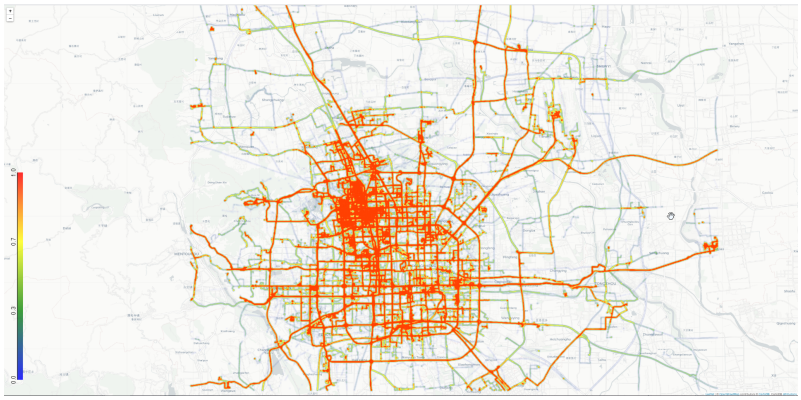
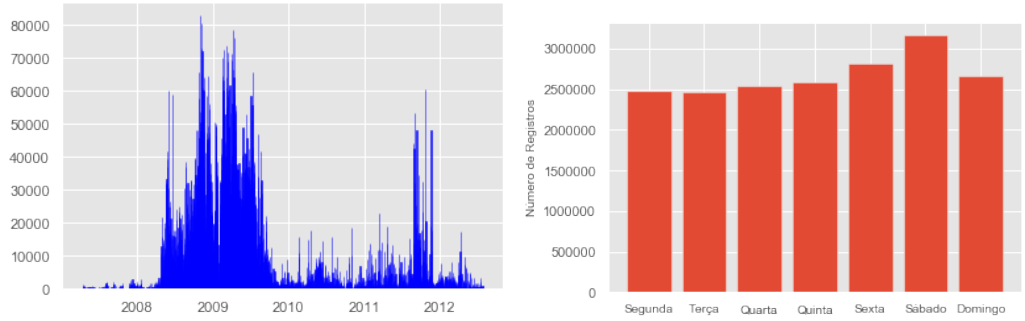


Figura 1. *Heat map* dos dados após filtragem.



(a) (b)

Figura 2. Número de registros por dia (a) e registros por dia da semana (b).

trabalho, utilizou-se a região de moradia dos usuários para a realização do agrupamento por comunidades.

No *dataset* utilizado não se possui a informação da localidade ou região de moradia de cada um dos usuários. Assim sendo, definiu-se a seguinte heurística para delimitar a região de moradia e, por consequência, a comunidade a que um usuário pertence:

- *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.*

O período T foi definido como sendo o intervalo entre 00:00 às 00:30 no horário local.

Por se tratar de uma heurística, existe a possibilidade de usuários que trabalham no horário de recolhimento e não residam na comunidade referente ao local de trabalho sejam tratados como membros dessas comunidades. Ademais, é possível também que usuários troquem de comunidade devido a mudanças de moradia ao longo de um determinado período de tempo, principalmente quando o período em que ocorreu a coleta de dados deste *dataset* é tão longo, aproximadamente 5 anos. Entretanto, como um usuário pode pertencer a mais de uma comunidade [Bess et al. 2002, Feng et al. 2015] e o modelo proposto atende a essa característica, esse comportamento foi mantido neste experimento. Dessa forma, os usuários serão agrupados em comunidades, podendo esses estar inclusos

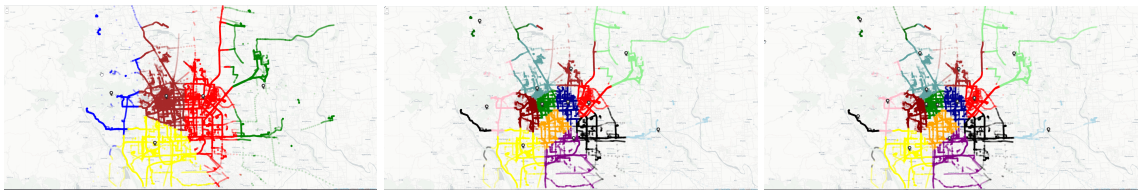
Segunda	Terça	Quarta	Quinta	Sexta	Sábado	Domingo
145	142	145	142	150	149	140

Tabela 1. Número de usuários ativos por dia da semana.

em mais de uma comunidade.

Para extrair as comunidades de cada usuário foi utilizado um algoritmo de clusterização. Como entradas para esse algoritmo foram utilizadas as coordenadas dos registros presentes no *dataset* que foram obtidas após o pré-processamento dos dados, detalhado anteriormente. Existem diversos algoritmos de clusterização que são utilizados na literatura, dentre os mais difundidos figuram o K-means e o DBSCAN. Neste artigo, o algoritmo de clusterização adotado foi o K-means. O K-means foi escolhido devido a sua performance em tempo de processamento, devido a quantidade de dados utilizados neste artigo. Como visto em [Ogbuabor and Ugwoke 2018], o DBSCAN tem um tempo de processamento cerca de 12 vezes maior que o do K-means para entregar um resultado com uma precisão similar.

A eficácia da clusterização utilizando o K-MEANS é diretamente relacionada a escolha do número de clusters K. Assim, é necessário a utilização de um método para a escolha do valor de K. Para isto, existem diversos métodos amplamente utilizados na literatura para se obter o melhor valor para K. Podemos citar o uso do Bayesian Information Criteria (BIC) por [Ferreira 2019], do Silhouette por [Nanjundan et al. 2019] e [Satre Meloy et al. 2019], do Elbow por [Nanjundan et al. 2019] e do Davies Bouldin por [Aggarwal et al. 2019]. Diante das várias de opções, utilizou-se o método Silhouette devido a sua ampla adoção, sua boa precisão, facilidade de uso quando comparado com os demais e compatibilidade com o tipo de dados utilizado neste experimento.



(a) (b) (c)

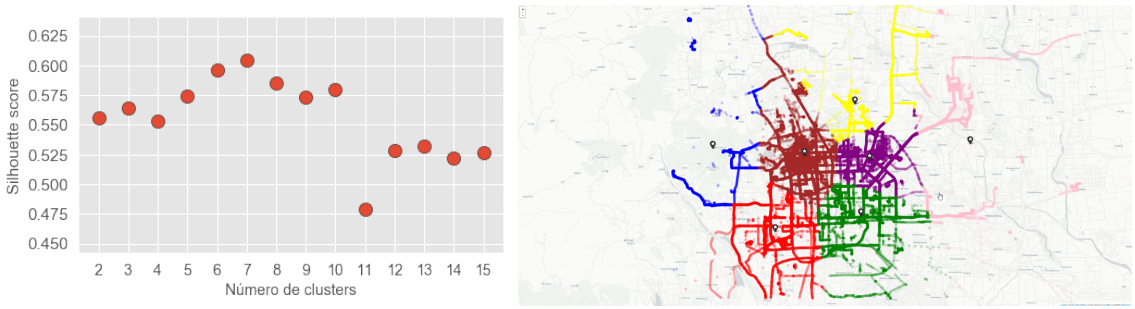
Figura 3. *Clusters* dos locais de residência das comunidades para K igual a 5, 11 e 15.

Para o cálculo do Silhouette Score, foi variado o parâmetro K de 2 a 15. Na Figura 3 são apresentados algumas das configurações de *clusters* obtidas.

Como detalhado em [Ogbuabor and Ugwoke 2018], o cálculo do *Silhouette Score* é assim obtido:

- 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 $S_j = \frac{(Y_j - X_j)}{Max(X_j, Y_j)}$

Os valores do *Silhouette score* para K entre 2 e 15 podem ser vistos na Figura 4(a). Vale notar que, quanto melhor o agrupamento, maior será o *Silhouette score*. Sendo assim, para as próximas etapas do experimento foi utilizado o K ótimo obtido pelo *Silhouette score* ($K = 7$) no K-MEANS. É apresentado na Figura 4(b) a distribuição geográfica, no mapa de Pequim, das áreas de residência de cada uma das comunidades geradas para K ótimo ($K = 7$). A apresentação desses dados se faz relevante para correlacionar as comunidades formadas com sua representação espacial real na cidade de Pequim.



(a) (b)

Figura 4. *Silhouette score* em função de K (a) e Distribuição espacial do local de residência das comunidades para K ótimo ($K = 7$) (b).

Além das sete comunidades formadas pelo K-means, foi criada uma oitava comunidade com todos os usuários que não possuíam registros noturnos e, portanto, não foram classificados em nenhuma das sete comunidades geradas. A distribuição dos usuários nas oito comunidades é apresentada na Tabela 2, 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 2. Número de usuários por comunidade.

4.3. Resultados

Após toda a configuração inicial tendo sido feita, o cálculo da influência de cada evento Ei foi realizado para cada comunidade Cj executando os passos a seguir:

1. Definiu-se a área de influência de cada evento E_i. Neste experimento definimos a área que estiver dentro do raio de 100 metros do epicentro do evento.
2. Calculou-se a distância geodésica de cada usuário U_k ativo, pertencente à comunidade C_j, ao centro do evento E_i analisado durante o período de ocorrência do evento. Optou-se pela distância geodésica sobre a distância Manhattan e Euclideana devido a sua maior precisão.
3. Realizou-se uma contagem do número de usuários, pertencentes à comunidade C_j, dentro do raio de influência do evento E_i, consolidando-se o número de usuários da comunidade C_j atingidos por tal evento.
4. Por fim, calculou-se a influência do evento Ei na comunidade C_j utilizando a seguinte equação:

$$influencia_{ij} = \frac{\sum_{k=0}^{n} U_k(j).atingido(i, U_k(j))}{\sum_{k=0}^{n} U_k(j)}$$

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

Conforme dito, foram analisados como certos eventos, organizados por dia da semana e localidade, afetam diferentes comunidades formadas com base no local de moradia. Foram selecionados onze locais para a ocorrência de eventos, com base na expectativa de um fluxo constante de pessoas de diferentes comunidades. Os pares (LAT, LON) desses locais estão ordenados abaixo, sendo eles:

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).

Os locais de ocorrência dos eventos listados acima, são demonstrados geograficamente na Figura 5. Nesta figura, os locais de evento são mostrados em cima do mapa de calor dos obtidos no dataset do posicionamento dos usuários. A partir da análise desta figura, pode-se visualizar de forma intuitiva, os locais que possuem uma quantidade maior ou menor de registros, e consequentemente, de movimentação de usuários.

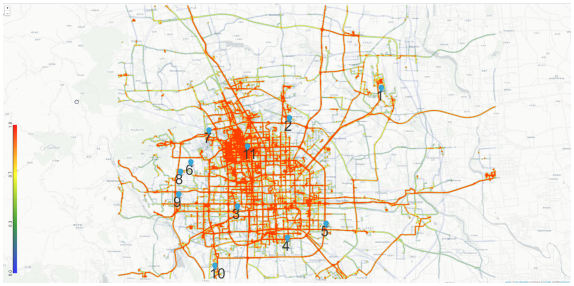


Figura 5. Localização geográfica dos locais de ocorrência dos eventos mostrado no *Heat Map* dos dados no mapa de Pequim.

Os resultados obtidos são apresentados nas Tabelas 3, 4, 5, 6, 7, 8 e 9, cada qual relacionada a um dia da semana. Os dados dessas tabelas estão organizadas na seguinte forma:

1. A quantidade de usuários únicos influenciados pelo evento Ei.
2. As colunas de "$C0$" à "$C7$" referem-se às oito comunidades identificadas. Os valores preenchidos em cada célula referem-se à influência relativa do evento no local i na comunidade Cj, equivalendo à fração de usuários de Cj afetados.
3. A coluna "Média_L" contém a média das influências relativas do local i para todas as comunidades Cj.
4. A linha "Média_C" contém a média das influências percebidas pela comunidade Cj para todos locais de evento i.

Vale destacar que os maiores valores obtidos para a influência do local i por comunidade foram marcados em negrito, exceto os valores obtidos para o Local 11, os quais foram destacados em itálico. Ressalta-se também que os resultados foram gerados a partir da análise de todos os dados presentes no *dataset* referentes ao enquadramento da região da grande Pequim.

Local	# Usuários	C1	C2	C3	C4	C5	C6	C7	C0	Média_L
01	08	07,4	09,8	13,0	09,8	11,1	10,2	**18,8**	00,0	10,1
02	**17**	15,7	25,5	30,4	19,5	**38,9**	25,4	34,4	00,0	23,8
03	16	12,0	15,7	17,4	**26,8**	16,7	15,3	15,6	03,6	15,4
04	15	07,4	15,7	17,4	17,1	16,7	**18,6**	12,5	01,8	13,4
05	11	07,4	15,7	**17,4**	14,6	16,7	13,6	15,6	03,6	13,1
06	03	00,9	02,0	00,0	**04,9**	02,8	01,7	00,0	01,8	01,8
07	12	08,3	13,7	**17,4**	14,6	16,7	11,9	06,3	03,6	11,6
08	06	02,8	03,9	**08,7**	07,3	05,6	03,4	00,0	03,6	04,5
09	13	10,2	05,9	13,0	**22,0**	08,3	08,5	09,4	01,8	09,9
10	10	09,3	07,8	08,7	**22,0**	05,6	10,2	12,5	00,0	09,5
11	*118*	*75,9*	*72,6*	*78,3*	*70,7*	*83,3*	*69,5*	*78,1*	*43,6*	*71,5*
Média_C	20,9	14,4	17,2	20,2	20,9	20,3	17,2	18,5	05,8	—

Tabela 3. Resultados de Segunda-Feira.

Local	# Usuários	C1	C2	C3	C4	C5	C6	C7	C0	Média_L
01	09	07,4	15,7	08,7	12,2	**19,4**	13,6	18,8	00,0	12,0
02	**14**	13,0	25,5	21,7	17,1	**33,3**	22,0	28,1	00,0	20,1
03	**14**	12,0	13,7	26,1	**31,7**	19,4	15,3	25,0	00,0	18,0
04	09	05,6	03,9	08,7	07,3	08,3	**10,2**	00,0	00,0	05,5
05	06	04,6	11,8	**17,4**	12,2	13,9	10,2	09,4	00,0	10,0
06	02	00,9	00,0	04,4	**04,9**	00,0	01,7	00,0	00,0	01,5
07	06	03,7	07,8	08,7	07,3	**11,1**	06,8	03,1	03,6	06,6
08	07	03,7	05,9	08,7	**09,8**	08,3	06,8	03,1	03,6	06,3
09	06	03,7	07,8	08,7	09,8	**11,1**	06,8	06,3	01,8	07,0
10	04	03,7	02,0	00,0	**09,8**	00,0	05,1	06,3	00,0	03,4
11	*112*	*71,3*	*70,6*	*78,3*	*70,7*	*77,8*	*71,2*	*75,0*	*43,6*	*69,9*
Média_C	17,2	11,8	15,0	17,4	17,6	18,5	15,5	16,0	04,8	—

Tabela 4. Resultados de Terça-Feira.

Ao analisar os resultados das tabelas, é possível observar que a influência dos eventos de qualquer local i em $C0$ é muito abaixo que para as demais comunidades, em todas as tabelas. Isso pode ser devido aos usuários de $C0$ não possuírem registros no período noturno, assim não há uma semântica clara de agrupamento desses usuários.

Nota-se também que o local 11 é o que gera as maiores influências para todas as comunidades em todas as tabelas, o que pode ser explicado pelo número de usuários ao redor desse local ser muito superior que o dos demais locais, em todos os dias da semana. Isso é coerente com o fato de se tratar de uma estação de trem muito movimentada.

Excluindo-se o local 11, verifica-se que a média das influências dos locais 2 e 3 são as próximas maiores em todos os dias da semana. O local 2 tem os melhores resultados na Segunda, Terça, Quinta e Domingo, enquanto que o local 3 possui os melhores resultados na Quarta, Sexta e Sábado. Vale notar que o local 3 também é uma estação de trem, porém com menos usuários monitorados passando por ela.

As comunidades $C3$, $C4$ e $C5$ são as mais influenciadas pelos locais de evento, o que também pode ser verificado pela média das influências. Nesse caso, $C3$ foi a mais influenciada na Quarta, Quinta, Sábado e Domingo. $C4$ foi a mais influenciada na Segunda e $C5$ na Terça e na Sexta.

Local	# Usuários	C1	C2	C3	C4	C5	C6	C7	C0	Média_L
01	08	07,4	11,8	17,4	14,6	16,7	11,9	**21,9**	00,0	12,7
02	16	13,9	23,5	30,4	19,5	**38,9**	22,0	31,3	00,0	22,5
03	**23**	20,4	27,5	**34,8**	34,2	25,0	27,1	25,0	00,0	24,3
04	12	07,4	07,8	08,7	**14,6**	08,3	11,9	09,4	00,0	08,6
05	05	03,7	07,8	**13,0**	07,3	11,1	06,8	09,4	01,8	07,7
06	05	01,9	00,0	**13,0**	09,8	00,0	01,7	00,0	01,8	03,6
07	14	12,0	19,6	**43,5**	14,6	27,8	17,0	21,9	01,8	19,8
08	07	03,7	03,9	**17,4**	12,2	05,6	05,1	03,1	01,8	06,6
09	06	04,6	05,8	**08,7**	07,3	08,3	05,1	03,1	01,8	05,6
10	09	08,3	07,8	13,0	**17,1**	05,6	10,2	09,4	00,0	09,0
11	*116*	*71,3*	*78,4*	*78,3*	*78,1*	*83,3*	*69,5*	*75,0*	*49,1*	*72,9*
Média_C	20,1	14,1	17,7	25,3	20,9	21,0	17,2	19,1	05,3	—

Tabela 5. Resultados de Quarta-Feira.

Local	# Usuários	C1	C2	C3	C4	C5	C6	C7	C0	Média_L
01	14	12,0	19,6	17,4	19,5	16,7	18,6	**31,3**	00,0	16,9
02	**18**	14,8	27,5	34,8	24,4	**38,9**	23,7	34,4	01,8	25,1
03	16	13,9	19,6	**34,8**	29,3	30,6	18,6	18,8	01,8	21,0
04	14	07,4	11,8	04,4	09,8	13,9	**15,3**	03,1	01,8	08,5
05	05	03,7	**07,8**	04,4	02,4	05,6	06,8	06,3	01,8	04,9
06	03	00,9	02,0	**08,7**	04,9	02,8	01,7	03,1	01,8	03,9
07	15	10,2	19,6	**30,4**	17,1	19,4	15,3	12,5	03,6	16,1
08	07	04,6	07,8	**13,0**	07,3	08,3	05,1	03,1	01,8	06,4
09	06	04,6	05,9	**13,0**	09,8	08,3	06,8	03,1	00,0	06,5
10	09	08,3	05,9	08,7	**14,6**	02,8	08,5	12,5	00,0	07,7
11	*110*	*72,2*	*66,7*	*73,9*	*73,2*	*80,6*	*69,5*	*68,8*	*40,0*	*68,2*
Média_C	19,8	13,9	17,7	22,2	19,3	20,8	17,3	18,0	05,0	—

Tabela 6. Resultados de Quinta-Feira.

Já $C7$ sofreu as maiores influências do local 01 em todos os dias da semana, a exceção de Terça. Esse dado contribui para a afirmação que o local 01 é de extrema relevância para a comunidade $C7$. Por outro lado, o local 6 é o que tem menor relevância geral para todas as comunidades, com média inferior a 4% em todos os dias da semana.

4.4. Cenário de aplicação

Nesta subseção é exemplificado um cenário hipotético no qual se poderia aplicar a abordagem proposta para identificar o impacto de eventos em um centro urbano.

O governo de Pequim poderia adotar uma política de vacinação destinada a comunidades com as características consideradas de risco para a contração de uma doença. Dada uma base de dados que contêm uma relação de custo horário de uso de um local para vacinação em relação ao dia da semana, com base nos resultados obtidos é possível modelar um problema de otimização para se conseguir a eficácia máxima de uma campanha de vacinação para uma comunidade específica k, dado um orçamento limite.

Supondo que essa eficácia é diretamente relacionada pelo número de horas de vacinação no local i no dia da semana j ponderado pela influência relativa desse par i, j,

Local	# Usuários	C1	C2	C3	C4	C5	C6	C7	C0	Média_L
01	09	07,4	07,8	04,4	14,6	08,3	11,9	**18,8**	00,0	09,2
02	17	14,8	25,5	26,1	22,0	**38,9**	23,7	28,1	00,0	22,4
03	**22**	18,5	27,5	**43,5**	41,5	36,1	28,8	25,0	00,0	27,7
04	13	07,4	11,8	08,7	14,6	13,9	**15,3**	06,3	00,0	09,8
05	13	12,0	17,7	**21,7**	12,2	19,4	17,0	12,5	00,0	14,1
06	02	01,9	**03,9**	00,0	02,4	02,8	03,4	03,1	00,0	02,2
07	18	15,7	25,5	**39,1**	22,0	33,3	22,0	21,9	01,8	22,7
08	07	05,6	11,8	13,0	09,8	**13,9**	10,8	06,3	01,8	09,1
09	07	06,5	05,9	04,4	**14,6**	08,3	08,5	09,4	00,0	07,2
10	10	08,3	03,9	04,4	**22,0**	00,0	06,8	15,6	00,0	07,7
11	*132*	*84,3*	*86,3*	*91,3*	*82,9*	*94,4*	*76,3*	*84,4*	*52,7*	*81,6*
Média_C	22,8	16,6	20,7	23,4	23,5	24,5	20,4	21,1	05,2	—

Tabela 7. Resultados de Sexta-Feira.

Local	# Usuários	C1	C2	C3	C4	C5	C6	C7	C0	Média_L
01	12	07,4	13,7	13,0	19,5	08,3	11,9	**25,0**	01,8	12,6
02	26	18,5	27,5	30,4	26,8	**47,2**	27,1	21,9	05,5	25,7
03	**31**	25,0	35,3	43,5	**43,9**	36,1	33,9	31,3	05,5	31,8
04	16	10,2	17,7	**26,1**	19,5	16,7	20,3	18,8	00,0	16,2
05	12	09,3	15,7	21,7	17,1	**22,2**	13,6	18,8	03,6	15,3
06	03	01,9	02,0	04,4	00,0	**05,6**	03,4	03,1	01,8	02,8
07	26	18,5	19,6	30,4	**31,7**	27,8	22,0	18,8	09,1	22,3
08	06	04,6	05,9	**08,7**	02,4	08,3	05,1	06,3	01,8	05,4
09	14	13,0	13,7	21,7	**22,0**	16,7	15,3	18,8	00,0	15,2
10	13	12,0	11,8	17,4	**26,8**	08,3	15,3	18,8	00,0	13,8
11	*108*	*75,9*	*72,6*	*82,6*	*80,5*	*86,1*	*74,6*	*65,6*	*34,6*	*71,6*
Média_C	24,3	17,9	21,4	27,3	26,4	25,8	22,1	22,5	05,8	—

Tabela 8. Resultados de Sábado.

temos a seguinte formulação:

$$f(k) = max(\sum_{i=1}^{11} \sum_{j=segunda}^{domingo} IR_{ij}(k).y_{ij})$$

sujeito a

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

$$y_{ij} \leq 16$$

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

onde $IR_{ij}(k)$ é a influência relativa do local i no dia j, cujos valores podem ser obtidos nas tabelas 3, 4, 5, 6, 7, 8 e 9. 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.

A título de exemplo reduzido, atribuímos valores hipotéticos ao problema de otimização definido acima considerando-se apenas o local 01, onde os custos e orçamento

Local	# Usuários	C1	C2	C3	C4	C5	C6	C7	C0	Média_L
01	11	10,2	17,7	17,4	17,1	19,4	17,0	**25,0**	00,0	15,5
02	**27**	21,3	43,1	47,8	31,7	**50,0**	33,9	40,6	03,6	34,1
03	19	17,6	25,5	**34,8**	31,7	25,0	22,0	31,3	00,0	23,5
04	10	06,5	07,8	04,4	09,8	11,1	**11,9**	06,3	00,0	07,3
05	14	10,2	19,6	21,7	12,2	**25,0**	17,0	21,9	03,6	16,4
06	02	01,9	02,0	**08,7**	02,4	02,8	01,7	03,1	00,0	02,9
07	14	12,0	17,7	21,7	19,5	**22,2**	15,3	15,6	01,8	15,8
08	02	01,9	03,9	**08,7**	02,4	05,6	03,4	03,1	00,0	03,7
09	07	06,5	05,9	08,7	**17,1**	08,3	06,8	09,4	00,0	07,9
10	07	06,5	03,9	04,4	**14,6**	05,6	06,8	09,4	00,0	06,4
11	*102*	*75,9*	*66,7*	*65,2*	*68,3*	*77,8*	*71,2*	*68,8*	*29,1*	*65,4*
Média_C	19,6	15,5	19,5	22,2	20,7	23,0	18,8	21,4	03,5	—

Tabela 9. Resultados de Domingo.

estão representados em milhares de Yuans para a comunidade $C1$. Para orçamento definiu-se 100 mil Yuans. Os demais valores, influência e custos para o local 01 por dia da semana, estão representados na Tabela 10.

#	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 10. Influência e Custos das ações por dia da semana.

Utilizando o método simplex para a formulação acima, foram obtidos os valores ótimos para o número de horas de vacinação no local 01 conforme o dia de semana j, cujos resultados são apresentados na Tabela 11. Esses valores indicam que, para a distribuição de custo horário do local 01 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 $C1$ dado o orçamento limite de 100 mil Yuans.

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

Tabela 11. Distribuição ótima de horas por dia da semana obtida através do modelo de otimização proposto.

5. Conclusão e Trabalhos Futuros

Este artigo propôs uma nova abordagem para a utilização de dados de mobilidade urbana voltada para a análise de eventos e suas influências em comunidades formadas por características de geolocalização de seus usuários. Essa nova abordagem permite uma análise mais aprofundada acerca das relações espaço-temporais entre essas comunidades e os locais de ocorrência de eventos, provendo suporte a tomadas de decisão voltadas para as mais diversas aplicações.

Para a validação da proposta, foi apresentado um estudo de caso baseado na cidade de Pequim voltado para uma campanha de vacinação com restrições orçamentárias. Nesse

estudo foi mostrado como gerar a combinação ótima entre locais de vacinação e duração dessas ações por dia da semana, objetivando a máxima influência dessa campanha em uma comunidade baseada em local de domicílio.

Como trabalhos futuros, deseja-se estender o estudo para outras semânticas de formação de comunidades, como por exemplo, comunidades formadas a partir da similaridade de mobilidade entre usuários proposta em [Ferreira 2019].

Referências

Aggarwal, S., Agarwal, N., and Jain, M. (2019). Performance analysis of uncertain k-means clustering algorithm using different distance metrics. In *Computational Intelligence: Theories, Applications and Future Directions - Vol I*, pages 237–245. Springer.

Bess, K. D., Fisher, A. T., Sonn, C. C., and Bishop, B. J. (2002). *Psychological Sense of Community: Theory, Research, and Application*, pages 3–22. Springer.

Feng, H., Tian, J., Wang, H. J., and Li, M. (2015). Personalized recommendations based on time-weighted overlapping community detection. *Information & Management*, 52(7):789 – 800. Novel applications of social media analytics.

Ferreira, D. L. (2019). *Characterization of Human Social Mobility Patterns Applied to Mobility Modelling and Opportunistic Networks*. PhD thesis, UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO.

Han, J., Pei, J., and Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Maze, T., Agarwai, M., and Burchett, G. (2006). Whether weather matters to traffic demand, traffic safety, and traffic operations and flow. *Transportation Research Record*, 1948:170–176.

Nanjundan, S., Sankaran, S., Arjun, C., and Anand, G. P. (2019). Identifying the number of clusters for k-means: A hypersphere density based approach. *arXiv preprint arXiv:1912.00643*.

Ogbuabor, G. and Ugwoke, F. (2018). Clustering algorithm for a healthcare dataset using silhouette score value. *International Journal of Computer Science & Information Technology*, 10(2):27–37.

Pregnolato, M., Ford, A., Wilkinson, S. M., and Dawson, R. J. (2017). The impact of flooding on road transport: A depth-disruption function. *Transportation Research Part D: Transport and Environment*, 55:67 – 81.

Satre Meloy, A., Diakonova, M., and Grunewald, P. (2019). What makes you peak? cluster analysis of household activities and electricity demand. European Council for an Energy Efficient Economy.

Stoltenberg, D., Maier, D., and Waldherr, A. (2019). Community detection in civil society online networks: Theoretical guide and empirical assessment. *Social Networks*, 59:120 – 133.

Zheng, Y., Fu, H., Xie, X., Ma, W.-Y., and Li, Q. (2011). *Geolife GPS trajectory dataset - User Guide*, geolife gps trajectories 1.1 edition. Geolife GPS trajectories 1.1.