Estudo do Comportamento de Consumo de Bebida em Centros Urbanos usando Redes de Sensoriamento Participativo

João A. S. Silva, Felipe D. Cunha, Silvio Jamil F. Guimarães

[1]Instituto de Ciências Exatas e Informática
Pontifícia Universidade Católica de Minas Gerais
R. Dom José Gaspar, 500 – Belo Horizonte – MG – Brazil

joao.silva.452811@sga.pucminas.br, {felipe,sjamil}@pucminas.br

Abstract. *The number of people affected by diseases related to the abuse of alcoholic beverages has grown considerably over the years, accounting for 3 million deaths per year worldwide. However, there are not many applications aimed at helping these people in recovery. Given this scenario, machine learning techniques can be found in the literature, which helps identify and characterize geographical regions conducive to alcohol consumption in large cities using urban data. This work analyzes the use of Location-Based Social Networks (LBSN) to assess the consumption of drinks in Tokyo and New York. Check-in data from bars and restaurants were collected, and, using machine learning techniques, it was possible to examine the drinking patterns of city residents. Results indicated that, although there were cultural differences in drinking habits between the two cities, users tended to consume more alcohol on weekends and at night. Additionally, it was possible to identify the regions most conducive to this consumption.*

Resumo. *O número de pessoas acometidas por doenças relacionadas ao abuso no consumo de bebidas alcoólicas tem crescido consideravelmente ao longo dos anos, contabilizando um total de 3 milhões de mortes ao ano em todo o mundo. Entretanto, não existem muitas aplicações voltadas à auxiliar essas pessoas em recuperação. Tendo em vista este cenário, na literatura podem ser encontradas técnicas de aprendizado de máquina que podem ajudar na identificação e caracterização de regiões geográficas propícias para o consumo alcoólico em grandes cidades utilizando dados urbanos. Este trabalho analisa o uso de Redes Sociais baseadas em Localização (LBSN) para avaliar o consumo de bebidas em Tóquio e Nova York. Foram coletados dados de check-ins em bares e restaurantes e, a partir de técnicas de aprendizagem de máquina, foi possível examinar os padrões de consumo de bebidas dos moradores das cidades. Resultados indicaram que, embora houvesse diferenças culturais nos hábitos de consumo de bebidas entre as duas cidades, os usuários tendiam a consumir mais álcool nos finais de semana e à noite. Além disso, foi possível identificar as regiões mais propícias a esse consumo.*

1. Introdução

Nos últimos anos houve um aumento significativo no número de doenças e mortes ocasionadas pelo consumo desordenado de bebidas alcoólicas, que é o principal causador de

200 doenças e de cerca de 3 milhões de mortes ao ano no mundo todo, o que corresponde a cerca de 5% das mortes globais [WHO 2022]. A partir desses dados, a compreensão sobre as causas e as consequências pode ser um fator chave para auxiliar às autoridades competentes a tomarem as medidas necessárias para o combate ao alcoolismo.

Atualmente, existem estudos acerca do combate a dependência alcoólica, como é feito em [Dulin et al. 2014, Gustafson et al. 2014], que propõem aplicações fornecem acompanhamento digital as pessoas que estão se recuperando dessa dependência através de funcionalidades projetadas para atender a este público. A funcionalidade que mais chama a atenção é a identificação de *High Risk Drinking Locations*, ou regiões vulneráveis, que consiste no cadastro dos locais onde o usuário tinha o hábito de consumir bebidas alcoólicas para que através da aplicação pudesse ser emitido um alerta para que o usuário ficasse atento nessas regiões. O mapeamento das *High Risk Drinking Locations* é feita a partir de um formulário dentro da aplicação de maneira estática e completamente dependente do usuário, o que levou ao principal objetivo do presente trabalho, a classificação dinâmica das regiões através de dados coletados de *check-ins* em *Location Based Social Networks (LBSNs)*, um dos principais sensores da computação urbana.

A computação urbana faz uso de uma grande quantidade de fontes de dados, como dispositivos de Internet das Coisas (IoT), dados de Redes Sociais Baseadas em Geolocalização (LBSN) e também dados estatísticos que facilitam na compreensão do ambiente urbano, assim conforme dito em [Rodrigues et al. 2019, Skora and Silva 2021, Silva et al. 2014b], a computação é capaz de fazer a diferença em diversas áreas e com o aumento na disponibilidade dos dados através de iniciativas para implementação de cidades inteligentes, surge a oportunidade de monitoração dos indivíduos em diversos aspectos, como por exemplo, na mobilidade desses indivíduos dentro da cidade, as suas rotinas, interesses, sentimentos, etc. Todos esses dados que podem ser coletados nos fornecem informações sobre diferentes domínios e assim como em [Machado et al. 2015] e [Gubert et al. 2022], é possível explorar dados de diferentes domínios através do sensoriamento em camadas e grafos multa aspecto, o que possibilita analisar a influência de fatores como o trânsito e as condições meteorológicas sob a mobilidade das pessoas de acordo com as classes sociais de uma cidade e a dinamicidade nos pontos de interesse.

A partir das aplicações propostas nos trabalhos [Le Falher et al. 2021, Dulin et al. 2014, Gustafson et al. 2014], o objetivo do presente trabalho é o desenvolvimento de uma aplicação capaz de mapear as regiões de cidades com base em suas atividades principais, com o foco principal em regiões que possuem alta probabilidade de consumo alcoólico a fim de fornecer dados úteis para o *marketing* relacionado a bebidas, logística para o definição de rotas de entrega, fiscalização dos órgãos competentes e para que pessoas em recuperação do alcoolismo possam evitar. A classificação dessas regiões é feita através de *check-ins* coletados das *LBSNs*, que depois de receberem os processamentos necessários, são submetidos a algoritmos de aprendizado de máquina para que seja feito o agrupamento dentro das cidades escolhidas a serem analisadas. Além da classificação dos *clusters* gerados, é possível entender mais sobre o fluxo das pessoas nos meios urbanos através dos sensores utilizados.

O trabalho está organizado da seguinte maneira, na Seção 2 são abordados os trabalhos que possuem maior relevância na motivação do trabalho e na metodologia escolhida. Na Seção 3 é discutida a metodologia com a descrição sobre a coleta de dados,

as bases de dados utilizadas, ferramentas e algoritmos de agrupamento. Na Seção 4 são abordados os resultados obtidos na classificação das regiões. Na Seção 5 são abordadas as possíveis aplicações para os resultados do trabalho e por fim, na Seção 6 é apresentada a conclusão do trabalho e a proposição de trabalhos futuros.

2. Trabalhos Relacionados

Diante do atual contexto global, de alta disponibilidade de dados coletados de sensores sociais, diversos autores têm apresentado soluções envolvendo a mobilidade urbana. Nesta Seção são apresentados alguns trabalhos que fazem uso desses dados para o entendimento do meio urbano.

Em [Zhang et al. 2021], os autores desenvolveram algoritmos voltados para um modelo de aprendizado conjunto multivisualização para incorporação de regiões urbanas utilizando grafos obtidos através de dados referentes a mobilidade humana dentro das cidades e atributos das regiões analisadas, que contabilizam *POI (Point of Interest)* e mobilidade humana dentro das regiões. A partir dos grafos gerados inicialmente, é aplicada a técnica de *Graph Attention Network (GAT)* para que o modelo aprenda sobre a representatividade dos vértices. Por fim, com os resultados já obtidos, foi criado um modelo de aprendizado conjunto voltado para permitir a colaboração de diferentes visualizações.

Já em [Dulin et al. 2014] e [Gustafson et al. 2014], os autores abordam a dificuldade de pacientes com distúrbio de uso de álcool em se manterem sóbrios após o início do tratamento para dar continuidade aos cuidados e para isso, propuseram duas aplicações, *StepAway* e *A-CHESS* respectivamente, que possuem como objetivo possibilitar que o usuário faça o acompanhamento do seu desempenho após o tratamento contra o alcoolismo. As aplicações propostas apresentam várias funcionalidades, mas a que mais chama atenção são as *High Risk Drinking Location*, em português, localizações com alto risco de consumo de álcool. A funcionalidade está presente em ambas aplicações propostas, mas dependem da interação do usuário, desta forma, definindo as *High Risk Drinking Locations* de forma que não acompanham a dinamicidade das cidades. O presente trabalho classifica as *High Risk Drinking Locations* de maneira dinâmica através de dados coletados em LBSNs.

Por fim, em [Le Falher et al. 2021], os autores buscam por meio de check-ins nas redes sociais estudar semelhanças entre vizinhanças a fim de fornecer ao usuário quais os bairros a serem visitados de acordo com a demanda fornecida, por exemplo, caso o usuário deseja fazer compras, a aplicação é capaz de direcioná-lo a uma vizinhança, caso queira jantar, será direcionado para outra vizinhança. A partir das inspirações obtidas em [Le Falher et al. 2021], o presente trabalho busca unir a funcionalidade de classificar regiões com a funcionalidade apresentada em [Dulin et al. 2014] e [Gustafson et al. 2014] para a identificação de *High Risk Drinking Location* dentro das cidades.

O presente trabalho se diferencia dos demais por fazer análises utilizando os *check-ins* de usuários, agrupados em uma base de dados própria, com o enfoque no sensoriamento de regiões que favorecem o consumo de bebida alcoólica. Assim, a partir desta análise, é possível a proposição de novas aplicações integrando a área da saúde com a área da computação urbana, assim como é proposto em [Dulin et al. 2014] e [Gustafson et al. 2014]. Esta integração pode ocorrer de forma dinâmica e não dependente dos usuários das aplicações a serem desenvolvidas. Para o desenvolvimento do trabalho são utilizados dois algoritmos de agrupamento para a formação das regiões a

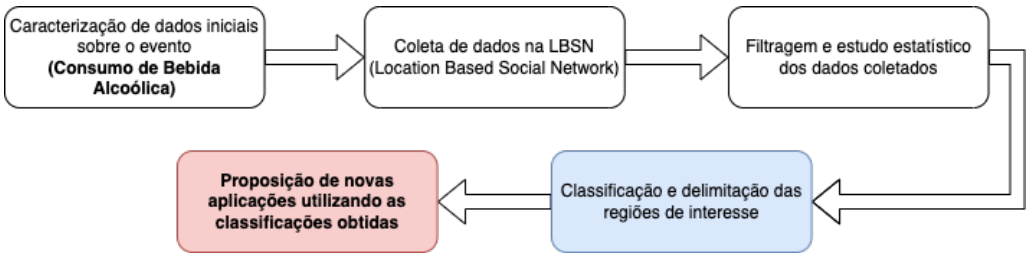


Figura 1. Fluxograma das Etapas do Trabalho.

serem analisadas, assim como executado em [Le Falher et al. 2021] e [Zhang et al. 2021]. Porém, neste trabalho foram utilizados os algoritmos *K-Means* e DBSCAN, que retornam os *clusters* a serem analisados e classificados posteriormente.

3. Metodologia

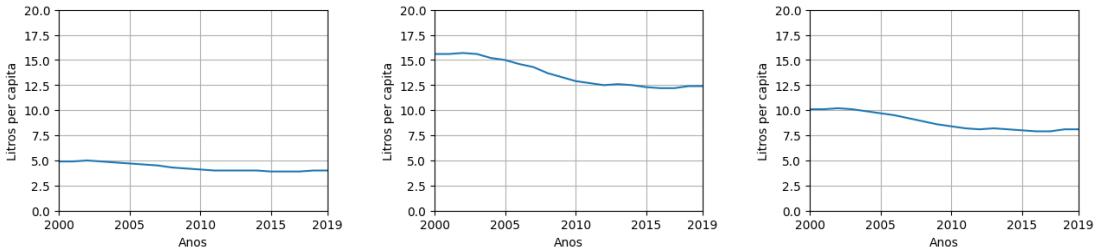
É ilustrada, na Figura 1, a metodologia para o estudo do comportamento de consumo de bebidas alcólicas, para tanto, a coleta de dados torna-se uma etapa fundamental do processo. São discutidas, nas próximas subseções, todas as etapas da metodologia que foram utilizadas neste trabalho.

3.1. Bases de dados

Foursquare é uma rede social mundialmente conhecida e permite com que seus usuários compartilhem a sua localização real e *feedbacks* a respeito dos locais com seu ciclo de amizades dentro da rede ou com seus seguidores em redes sociais vinculadas, como o *Twitter* e o *Facebook*. Com o passar do tempo, o *Foursquare* migrou a funcionalidade de *check-ins* para uma outra aplicação da empresa dedicada para essa funcionalidade específica, chamado *Swarm*, que possibilita a mesma interação que os usuários possuíam dentro do próprio *Foursquare*, mas agora com uma plataforma totalmente dedicada a esse meio. De acordo com a política de privacidade do *Foursquare*, os *check-ins* são considerados como informação privada, mas alguns usuários optam por compartilhar a sua localização via *Twitter*, o que faz com que os *check-ins* se tornem públicos e possibilita o acesso ao dado sem violar as regras de nenhuma das redes sociais utilizadas, conforme considerado em [Silva et al. 2014a]. A partir da obtenção dos *check-ins* pela *Twitter API*, foram coletadas aproximadamente 2.7 milhões de instâncias entre Maio de 2022 e Janeiro de 2023 ao redor do mundo inteiro, possibilitando uma análise mais abrangente sobre o comportamento de cada país.

Para análise e entendimento do consumo de bebidas alcoólicas, foi utilizada uma base de dados disponibilizada no domínio da *World Health Organization*(WHO[1]), por meio do programa *The Global Health Observatory* que disponibiliza dados do *Sustainable Development Goals* (SDG[2]) que aborda 17 objetivos definidos pelas nações parceiras da Organização das Nações Unidas (ONU), a fim de abdicar a pobreza e a desigualdade, melhorar a saúde, justiça e prosperidade, além de melhorar a conservação do planeta. O indicador utilizado para o entendimento sobre o consumo de bebidas alcoólicas no mundo foi o *Indicator* 3.5.2, que está ligado ao objetivo 3 do SDG, que busca garantir saúde e bem estar a todas as idades, alvo 5, que busca fortalecer a prevenção e tratamento do abuso

[1]https://who.int/
[2]https://sdgs.un.org/



(a) Feminino (b) Masculino (c) Ambos

Figura 2. Consumo de álcool em litros per capita (15+) por ano no Japão.

de substâncias e indicador 2, que mostra a quantidade em litros de puro álcool consumidos ao longo dos anos por pessoas maiores de 15 anos em vários países.

Os gráficos ilustrados na Figura 2 representam dados referentes ao *Indicator* 3.5.2 para o Japão. Infelizmente, a previsão para os próximos anos não é uma tarefa fácil de ser obtida uma vez que não observou-se padrão adequado para tal previsão. Cumpre reforçar que todos os gráficos para todos os países comportam-se de forma similar. Além disto, tal previsão possivelmente apresentaria grandes diferenças devido ao contexto mundial da pandemia de *COVID-19*. A base escolhida dentro do indicador explicado anteriormente apresenta o volume de álcool consumido em 186 países, entre 2000 e 2019. Contabilizando pessoas com mais de 15 anos de idade e dividido por gênero, mostrando a diferença do consumo entre homens e mulheres e a média de ambos.

3.2. Ferramentas

Neste trabalho, foram utilizadas diversas plataformas para o desenvolvimento de ferramentas necessárias para cada etapa do estudo do comportamento de consumo. Para a coleta dos *check-ins* foi utilizada a *Twitter API*, onde foi utilizada a licença gratuita cuja permite a coleta de 500.000 *tweets* por mês. A coleta foi executada em uma máquina virtual hospedada na *Microsoft Azure* por um código escrito em *Python*, linguagem de programação que também foi utilizada para o tratamento dos *check-ins* coletados, para o pré-processamento dos dados e para a análise de todos os dados dentro das plataformas *Jupyter Notebook* e *Google Colab*, onde foram executados os algoritmos de agrupamento para a caracterização das regiões. Todos os códigos utilizados no trabalho estão disponíveis em um repositório no *GitHub* [3].

3.3. Pré-processamento

Considerando que a base de dados contêm mais informações que àquelas necessárias para nossa análise, tornou-se necessária a remoção de informações irrelevantes dos *tweets* coletados. Para tanto, filtros foram desenvolvidos para o tratamentos destes dados. Para a obtenção das informações necessárias em cada *check-in* coletado, foi necessário a abertura do código HTML referente ao *link* do *check-in* no *Swarm* por meio de uma aplicação desenvolvida na linguagem de programação *Python*. Por meio do tratamento feito são retornadas as informações desejadas, o que descartou o uso da *Places API*, plataforma da *Foursquare*. Após a etapa de processamento de dados, foi obtida a base de dados inicial, conforme descrita na Tabela 1, contendo aproximadamente 2.7 milhões

[3]https://github.com/joaoaugustoss/Consumo-Alcool

Atributo	Descrição do Atributo
venueID	Identificador da *venue* no *Foursquare*
userID	Identificador do usuário no *Swarm*
venueName	Nome da *venue* onde foi efetuado o *check-in*
category	Categoria da *venue* onde foi efetuado o *check-in*
country	País da *venue* onde foi efetuado o *check-in*
city	Cidade da *venue* onde foi efetuado o *check-in*
timestamp	Horário no qual o *check-in* foi compartilhado no *Twitter*
latitude	Latitude da *venue* onde foi efetuado o *check-in*
longitude	Longitude da *venue* onde foi efetuado o *check-in*

Tabela 1. Descrição da base de dados utilizada no trabalho.

de instâncias, entretanto, graças à remoção das linhas com dados ausentes, a base filtrada contêm aproximadamente 1 milhão instâncias, como ilustrado na Tabela 2.

3.4. Agrupamento

A fim de gerar *clusters* para classificar as cidades com base em regiões, foram utilizados dois algoritmos para o agrupamento de dados, sendo eles o *DBSCAN*, algoritmo baseado na densidade e o *K-Means*. A escolha dos algoritmos se deu devido a capacidade de ambos em trabalharem com dados georeferenciados e pela simplicidade em sua implementação.

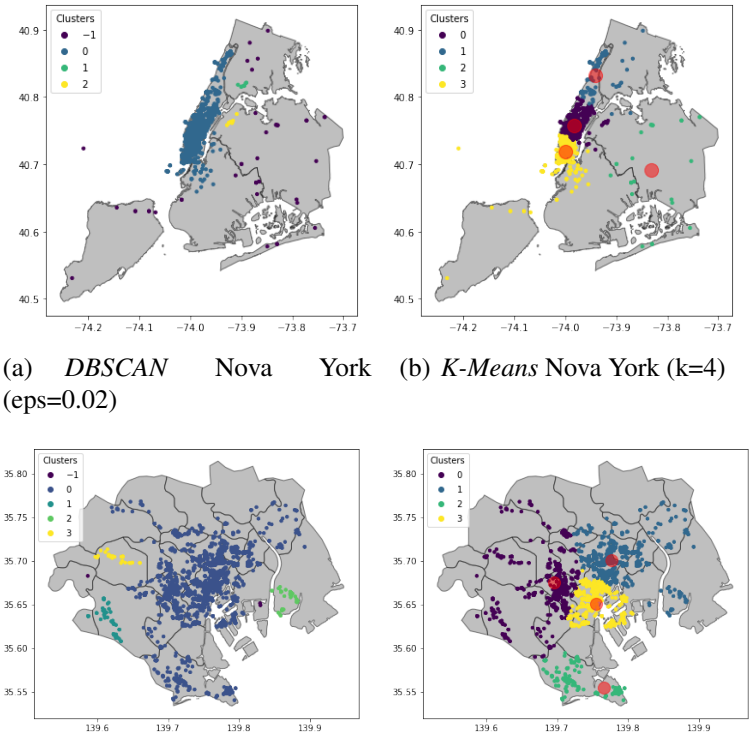
Foram executados os algoritmos de agrupamento em dois momentos. No primeiro momento foi considerado apenas dados referentes à geolocalização das instâncias, que obteve o melhor resultado e consequentemente sendo utilizado para as análises expostas na Seção 4. Um exemplo deste agrupamento utilizando apenas coordenadas geográficas pode ser visto na Figura 3. No segundo momento, foi agregada a classificação das *venues* a serem agrupadas com os seus dados geográficos. Após a execução, nenhum dos algoritmos executados obtiveram bons resultados, já que com a agregação das categorias nos dados a serem agrupados, os algoritmos passaram a focar nos dados referentes às categorias e não nos dados georreferenciais, o que levou a sobreposição de *clusters* em ambos os algoritmos executados. Tendo em vista o contraste nos resultados, foi considerado para a análise apenas o agrupamento advindo dos algoritmos de agrupamento considerando apenas as coordenadas geográficas das localidades.

3.4.1. DBSCAN

DBSCAN, que significa *Density-Based Spatial Clustering of Applications with Noise*, é um algoritmo de aprendizado de máquina voltado para o agrupamento baseado em densidade, avaliando a distância entre os pontos a partir de um ponto inicial aleatório para realizar a

Cidade	*Ckeck-ins*	*Venues*
Tóquio	17,320	2,636
Nova York	1,916	1,017
Outras	966,104	230,511
Total	985,661	234,164

Tabela 2. *Check-ins* coletados por cidade após processamento dos dados.



(a) *DBSCAN* Nova York (eps=0.02)

(b) *K-Means* Nova York (k=4)

(c) *DBSCAN* Tóquio (eps=0.02)

(d) *K-Means* Tóquio (k=4)

Figura 3. Comparativo do agrupamento obtido pelo K-means e DBSCAN.

diferenciação entre os grupos encontrados, o que possibilita com que o próprio algoritmo determine o número de *clusters* a serem formados após a sua execução. O algoritmo proposto em [Ester et al. 1996] possui lados negativos, pois dependendo da escolha do ponto inicial podem ser gerados *clusters* diferentes e além do fato que o *DBSCAN* não trabalha bem com grandes conjuntos de dados, o que torna inviável a sua escalabilidade. Em contrapartida, apresenta excelentes resultados classificando os *outliers* como ruídos, descartando a necessidade de de sua eliminação no pré-processamento dos dados.

O algoritmo *DBSCAN* foi aplicado nas duas cidades escolhidas para análise levando em consideração apenas as coordenadas geográficas dos dados. Os resultados obtidos não foram adequados, mesmo após o ajuste do hiperparâmetro 'eps', que define a distância necessária entre dois pontos para que sejam considerados do mesmo grupo. Para o agrupamento dos dados referentes à cidade de Tóquio, como ilustrado na Figura 3(c), foi necessário definir o 'eps' com o valor de 0.02, o que resultou em um agrupamento com 4 *clusters* e 1 grupo contendo 8 instâncias classificadas como ruído. Já na cidade de Nova York, também foi utilizado o valor 0.02 no hiperparâmetro 'eps', o que gerou um agrupamento com 2 *clusters* e 1 grupo contendo 50 instâncias classificadas como ruído, assim como é ilustrado na Figura 3(a). Os grupos classificados como ruído se encontram no *cluster* -1 nas Figuras 3(a) e 3(c).

3.4.2. K-Means

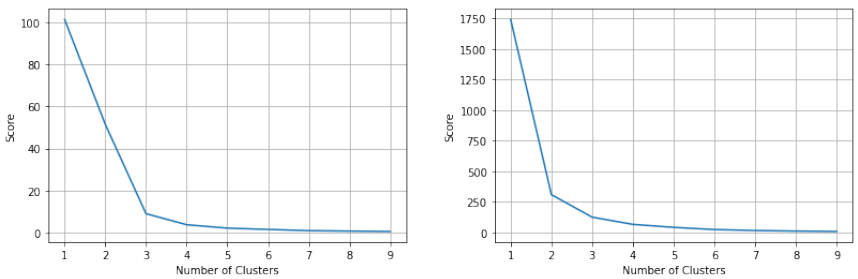
K-Means é um algoritmo de aprendizado de máquina não-supervisionado que foi projetado para o agrupamento de dados levando em conta as suas características. Ao contrário do

algoritmo *DBSCAN*, para aplicação do *K-Means* é necessário a definição prévia do número de *clusters* que é desejado e para a definição desse número foi utilizado o método *Elbow* representado na Figura 4. O *Elbow* tem a função de testar a variância dos dados em relação ao número de *clusters* a fim de retornar o número ideal de *clusters* para o agrupamento a ser executado. Na Figura 4(b), é possível identificar o ponto de inflexão no gráfico próximo ao valor 3 no eixo x, o que indica que a partir deste valor não haverá ganho no agrupamento com o aumento do número de *clusters*. A mesma análise pode ser feita na Figura 4(a) que se refere ao *Elbow* para a cidade de Nova York, onde o ponto de inflexão se aproxima do valor 2 no eixo x.

Após a separação das coordenadas geográficas para a execução do algoritmo *K-Means*, foram obtidos resultados soberanos aos gerados pelo *DBSCAN*, com a diferença na classificação dos *clusters* com maior concentração de *check-ins* que se encontram no centro das cidades analisadas, assim como pode ser visto na Figura 3, o que demonstra a dificuldade apresentada pelo *DBSCAN* no agrupamento de dados muito densos. Para a definição do número de *clusters*, foi definido o valor de k=4 para o agrupamento realizado nas instâncias referentes à cidade de Tóquio, assim como é mostrado na Figura 3(d) e o resultado obtido foi visualmente diferente em relação ao *DBSCAN*, que também agrupou os dados em 4 *clusters*. Na execução com os dados referentes a cidade de Nova York, representado na Figura 3(b), foi utilizado o valor de k=4. Ao comparar os resultados, novamente o algoritmo *K-Means* se mostrou superior no agrupamento, onde foi capaz de separar a região central em grupos distintos e os pontos mais isolados, mostrando a real diferença na implementação dos algoritmos e justificando a escolha do agrupamento obtido por meio do algoritmo *K-Means* para a análise das regiões.

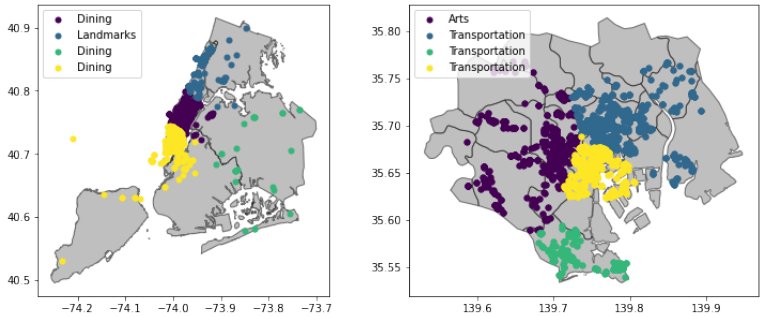
4. Resultados

Nesta Seção são avaliados os resultados das classificações das vizinhanças, que foram obtidas a partir dos *clusters* gerados a partir do algoritmo *K-Means*, que foi exemplificado na Seção 3.4. Junto a isso, dos 18.7 mil dados iniciais após a execução da *clusterização* com os dados referentes às cidades escolhidas para análise, em Nova York, obtivemos 1.9 mil *check-ins* em 1.1 mil *venues* únicas divididas em 4 regiões. Já em Tóquio, foram obtidos 17.3 mil *check-ins* em 2.6 mil *venues* únicas divididas também em 4 regiões. A análise dos dados obtidos após o agrupamento foi dividida em dois momentos, a análise espacial, onde foram considerados apenas dados georreferenciais referentes à publicação do *check-in* para a classificação do *cluster*. Esta análise se encontra na Seção 4.1. Já para a análise



(a) Nova York (b) Tóquio

Figura 4. Gráfico método *Elbow* para as cidades de Nova York e Tóquio.



(a) Nova York (b) Tóquio

Figura 5. Classificação Espacial de *check-ins* pra Nova York e Tóquio.

temporal, foram consideradas informações referentes ao dia e ao horário de publicação dos *check-ins*, onde houve a divisão entre dias de semana e faixas de horário. Esta análise pode ser encontrada na Seção 4.2. A divisão das análises favorece ao entendimento da mobilidade humana nas regiões analisadas, tornando possível a compreensão na diferença do comportamento humano conforme o dia da semana e o período do dia.

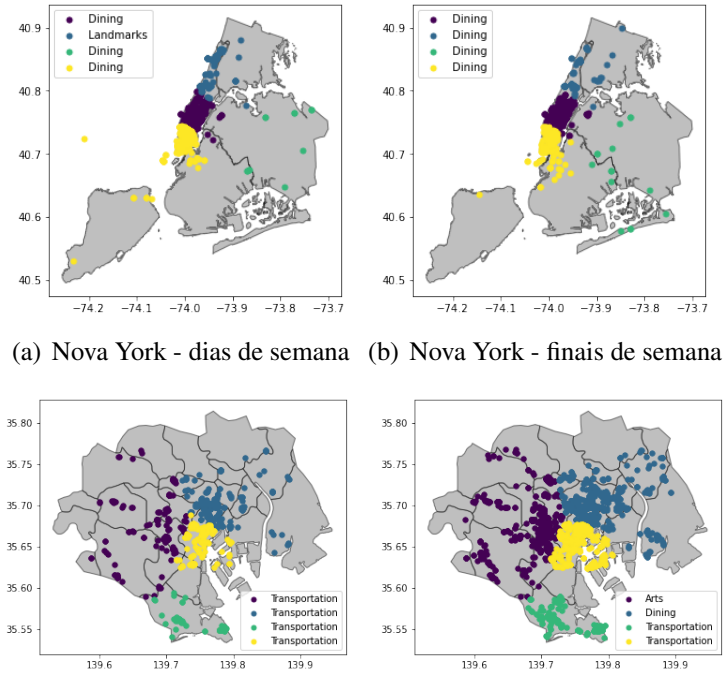
4.1. Análise Espacial

Após a delimitação das regiões gerada pelo *K-Means*, foi feita a classificação das regiões obtidas. Nesta primeira análise, foi contabilizado o número de *check-ins* em cada *venue* dentro de um determinado *cluster*. Após a obtenção do número de *check-ins* em cada *venue*, foi analisada a categoria com o maior número de *check-ins* e a partir desta categoria foi gerada a classificação do *cluster*, podendo ser *Entertainment, Business, Community, Dining, Event, Health, Landmarks, Retail, Sports, Transportation* ou *Residence*, que correspondem às categorias encontradas no *Foursquare* referentes a cada *venue*.

Após a análise dos dados, foi realizada a classificação das regiões como mostrado na Figura 5. Notou-se a dominância da classificação *Dining* na cidade de Nova York (ver Figura 5(a)) e a dominância da classificação *Transportation* na cidade de Tóquio (ver Figura 5(b)), entratanto, estas classificações não parecem ser precisa acerca das cidades. Para a identificação de regiões com alto consumo de álcool, foram obtidos resultados semelhantes para a classificação *Dining*, mostrando a dominância de restaurantes em todas as regiões. A partir destes resultados pode-se arguir se as pessoas realmente buscam essas categorias a todo momento dentro da alta dinamicidade presente nas cidades que foram analisadas.

4.2. Análise Temporal

A partir da análise espacial, é possível afirmar que o dia e a hora no qual o usuário efetua o seu *check-in* possui grande relevância na classificação da região procurada pelos usuários, podendo definir pontos de interesse momentâneos, como eventos em cidades, sendo eles eventos profissionais que geralmente ocorrem durante a semana e durante o horário comercial ou eventos relacionados ao lazer das pessoas, como por exemplo shows que em sua normalidade ocorrem durante o final de semana e durante a noite. A detecção desses pontos de interesse é realizada pela alta concentração de *check-ins* em um curto intervalo de tempo em uma mesma *venue*, ou em coordenadas geográficas próximas



(a) Nova York - dias de semana (b) Nova York - finais de semana

(c) Tóquio - dias de semana (d) Tóquio - finais de semana

Figura 6. Classificação de regiões considerando os dias da semana.

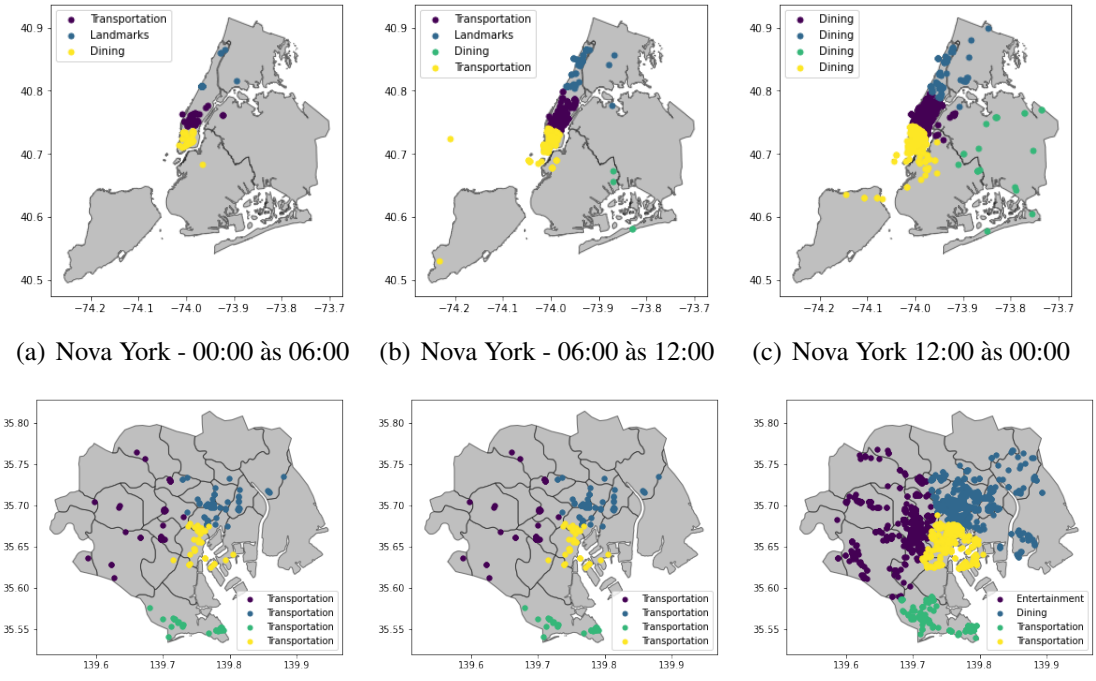
umas das outras, assim como dito em [Silva et al. 2013]. A análise considerando dados temporais referentes ao dia e a hora do *check-in* é dividida em dois momentos, a análise considerando o dia da semana do dado na Seção 4.2.1 e a análise com os dados referentes ao horário na Seção 4.2.2.

4.2.1. Dias de Semana

Para a análise e classificação das regiões com base nos dias que foram registrados os *check-ins* dos usuários, foram divididas as informações temporais em grupos, em que o primeiro grupo corresponde aos registros feitos em dia da semana (de segunda à sexta) e o segundo grupo corresponde aos finais de semana (sábado e domingo).

A partir dos dados ilustrados na Figura 6, foi possível visualizar pequenas diferenças entre as classificações realizadas nas cidades de Nova York, Figuras 6(a) e 6(b) e Tóquio, Figuras 6(c) e 6(d). Na cidade de Nova York, a classificação *Dining* continuou predominante em meio a todas as regiões classificadas durante a semana e também durante os finais de semana. Já na cidade de Tóquio, a classificação *Transportation* também prevaleceu diante das demais classificações possíveis.

Partindo para a análise dos *check-ins* vinculados a categoria *Dining* para a identificação de regiões passíveis para o consumo alcoólico, também foi possível observar a falta de dinamicidade nas classificações, já que *venues* consideradas *Restaurant* foram mais uma vez predominantes nas classificações das regiões. Após a análise destes resultados, também foi possível concluir que a divisão entre dias de semana e finais de semana não é capaz de retratar o fluxo de pessoas dentro das cidades, o que levou a acreditar na necessidade de uma análise com base no horário cujo *check-in* foi coletado.



(a) Nova York - 00:00 às 06:00 (b) Nova York - 06:00 às 12:00 (c) Nova York 12:00 às 00:00

(d) Tóquio - 00:00 às 06:00 (e) Tóquio - 06:00 às 12:00 (f) Tóquio - 12:00 às 00:00

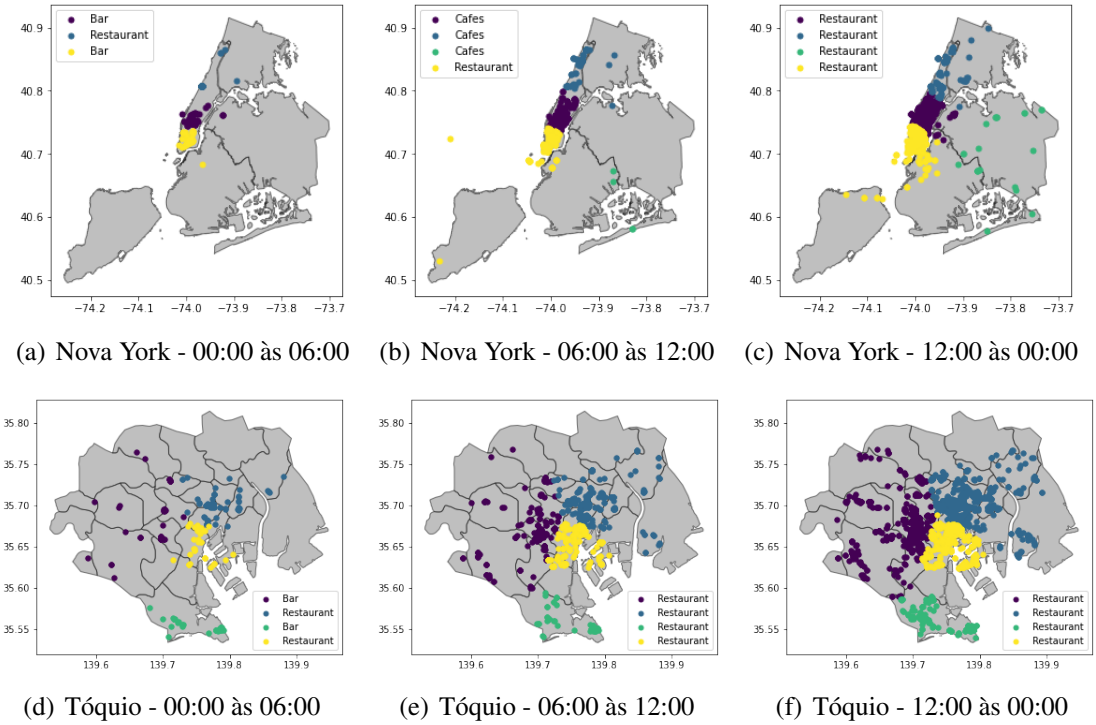
Figura 7. Classificação de regiões considerando o horário que os *check-ins* foram realizados.

4.2.2. Horário

Além da detecção de pontos de interesse, a análise temporal também é capaz de descrever o estilo de vida das cidades e o seu comportamento. Nas Figuras 7(d) e 7(e) é possível observar a dominância de *check-ins* classificados como *Transportation* entre 00:00 e 12:00 na cidade de Tóquio, Figuras 7(d) e 7(e), demostrando a alta mobilidade das pessoas durante essa faixa de horário. Também é possível notar a dominância de *check-ins* classificados como *Transportation* nas regiões 2 e 3 na faixa de horário entre 12:00 e 00:00, Figura 7(f), o que pode indicar a presença de terminais rodoviários e/ou ferroviários nas regiões, justificando a constância na classificação da região durante todo o período analisado.

Assim como em Tóquio, representada pela Figura 7(e), também é possível visualizar a concentração de *check-ins* voltados aos meios de transporte nos horários entre às 06:00 e 12:00 na cidade de Nova York, exibida na Figura 7(b). As Figuras 7(f) e 7(c) mostram a dominância da categoria *Dining* de 12:00 até às 00:00 em todas as regiões analisadas nas duas cidades, o que desperta o interesse em passar para a análise referente apenas aos dados categorizados como *Dining* que dominam metade do dia em duas cidades que apresentam culturas muito distintas umas das outras.

Analisando apenas os *check-ins* categorizados como *Dining*, que podem ser visualizados na Figura 8, é possível observar a diferença nas buscas feitas pelas pessoas em relação ao tipo de local que procuraram na hora de se alimentarem. O que mostra a presença elevada da categoria *Bar* entre $00:00$ e $06:00$, Figuras 8(a) e 8(d) em ambas cidades analisadas, a dominância de *Cafes* na cidade de Nova York entre $06:00$ e $12:00$,



(a) Nova York - 00:00 às 06:00 (b) Nova York - 06:00 às 12:00 (c) Nova York - 12:00 às 00:00

(d) Tóquio - 00:00 às 06:00 (e) Tóquio - 06:00 às 12:00 (f) Tóquio - 12:00 às 00:00

Figura 8. Classificação da categoria *Dining* do local, considerando o horário em que os *check-ins* foram realizados.

Figura 8(b), e a dominância de *Restaurant* ao longo do restante do dia tanto em Tóquio quanto em Nova York, Figuras 8(c), 8(e) e 8(f).

As classificações que levam em conta o horário dos *check-ins* são as que obtiveram maior sucesso em relação às classificações localizadas nas Seções 4.1 e 4.2.1 devido a variância visível nas classificações, o que é condizente com a mobilidade constante de pessoas e a constante mudança em seus interesses de acordo com o horário nas cidades analisadas que possuem fluxo de pessoas elevado durante todo o dia.

5. Aplicações

Com base nas classificações apresentadas na Seção 4 e com os agrupamentos realizado na Seção 3.4, é possível a proposição de aplicações nas seguintes vertentes:

Fiscalização: Direcionamento de órgãos públicos responsáveis pela fiscalização de trânsito e de regulação de substâncias não permitidas para menores de idade para áreas classificadas como *Dining* para a prevenção de possíveis crimes.

***E-Health*:** Identificação das regiões que podem apresentar o consumo de álcool, apresentando assim possível risco para pessoas que estão se recuperando do vício alcoólico, a fim de prevenir a reincidência no vício.

Logística: Auxílio no planejamento da logística de entrega de suprimentos destinados a estabelecimentos voltados a alimentação e bebidas.

***Marketing*:** Mapeamento de regiões onde empresas e comércios obterão maior sucesso ao direcionar os seus serviços de *marketing*.

Tráfego de veículos: Planejamento de rotas de tráfego de veículos a fim de evitar regiões com alta concentração de *check-ins* que podem sinalizar possíveis congestionamentos.

6. Conclusão e Trabalhos Futuros

Tendo em vista os resultados apresentados na Seção 4 sobre o consumo de bebidas alcoólicas, métodos devem ser explorados para auxiliarem indivíduos que precisam evitar as bebidas alcoólicas também para auxiliar na fiscalização, marketing e logística para regiões que apresentam grande concentração de bares, restaurantes e outras localidades que comercializam insumos alcoólicos. Sendo assim, este trabalho realizou a análise de *check-ins* coletados por meio de *LBSNs* por meio de algoritmos de agrupamento para detectar *clusters* com base na sua geolocalização, em seguida, os *clusters* foram classificados quanto à categoria dos locais com maior popularidade, que foi medida com base na hora, no dia e também desconsiderando informações temporais.

Foi explorado o uso de técnicas de agrupamento para a classificação das regiões. O uso do aprendizado de máquina possibilitou a rotulação dos grupos gerados e a classificação desses grupos quanto a categoria das *venues* mais frequentes em determinados grupos. A partir dessas classificações, foi possível a identificação das *High Risk Drinking Location* por meio dos *check-ins* concentrados no período da noite e nos finais de semana, assim como foi proposto inicialmente para o trabalho, o que possibilita a compreensão da mobilidade dentro das cidades analisadas e a visualização de que mesmo em meio a culturas muito distintas dentro das cidades de Nova York e Tóquio, os sensores sociais são capazes de retratar o que acontece dentro dos centros urbanos e mapear o consumo de bebidas alcoólicas de acordo com o dia e a hora que se deseja analisar.

Para trabalhos futuros, torna-se fundamental explorar a parametrização de forma exaustiva dos algoritmos de agrupamentos utilizado, além disto, estudar diferentes técnicas para o agrupamento dos *check-ins* e a implementação de uma das possíveis aplicações sugeridas na Seção 5 que seja capaz de utilizar as classificações obtidas acerca das regiões.

Agradecimentos

Os autores agradecem ao Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq – (PQ 310075/2019-0), à Fundação de Amparo à Pesquisa do Estado de Minas Gerais – FAPEMIG – (Projeto PPM-00006-18 e PIBIC 2022/28009), e à Pontifícia Universidade Católica de Minas Gerais (PUC Minas) pelo suporte para o desenvolvimento deste trabalho.

Referências

Dulin, P. L., Gonzalez, V. M., and Campbell, K. (2014). Results of a pilot test of a self-administered smartphone-based treatment system for alcohol use disorders: usability and early outcomes. *Substance abuse*, 35(2):168–175.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231.

Gubert, F. R., Munaretto, A., and Silva, T. H. (2022). Multilayered analysis of urban mobility. In *Anais Estendidos do XXVIII Simpósio Brasileiro de Sistemas Multimídia e Web*, pages 57–60. SBC.

Gustafson, D. H., McTavish, F. M., Chih, M.-Y., Atwood, A. K., Johnson, R. A., Boyle, M. G., Levy, M. S., Driscoll, H., Chisholm, S. M., Dillenburg, L., et al. (2014). A smartphone application to support recovery from alcoholism: a randomized clinical trial. *JAMA psychiatry*, 71(5):566–572.

Le Falher, G., Gionis, A., and Mathioudakis, M. (2021). Where is the soho of rome? measures and algorithms for finding similar neighborhoods in cities. *Proceedings of the International AAAI Conference on Web and Social Media*, 9(1):228–237.

Machado, K., Silva, T. H., de Melo, P. O. V., Cerqueira, E., and Loureiro, A. A. (2015). Urban mobility sensing analysis through a layered sensing approach. In *2015 IEEE International Conference on Mobile Services*, pages 306–312. IEEE.

Rodrigues, D. O., Santos, F. A., Akabane, A. T., Cabral, R., Immich, R., Junior, W. L., Cunha, F. D., Guidoni, D. L., Silva, T. H., Rosário, D., et al. (2019). Computação urbana da teoria à prática: Fundamentos, aplicações e desafios. *arXiv preprint arXiv:1912.05662*.

Silva, T., De Melo, P. V., Almeida, J., Musolesi, M., and Loureiro, A. (2014a). You are what you eat (and drink): Identifying cultural boundaries by analyzing food and drink habits in foursquare. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8, pages 466–475.

Silva, T. H., De Melo, P. O. V., Almeida, J. M., Salles, J., and Loureiro, A. A. (2013). A picture of instagram is worth more than a thousand words: Workload characterization and application. In *2013 IEEE International Conference on Distributed Computing in Sensor Systems*, pages 123–132. IEEE.

Silva, T. H., Vaz de Melo, P. O., Almeida, J. M., Salles, J., and Loureiro, A. A. (2014b). Revealing the city that we cannot see. *ACM Transactions on Internet Technology (TOIT)*, 14(4):1–23.

Skora, L. E. B. and Silva, T. H. (2021). Comparing international movements of tourists: Official census versus social media. In *Anais Estendidos do XXVII Simpósio Brasileiro de Sistemas Multimídia e Web*, pages 45–48. SBC.

WHO (2022). World health organization. `https://www.who.int/news-room/fact-sheets/detail/alcohol`. Accessed: 2022-09-30.

Zhang, M., Li, T., Li, Y., and Hui, P. (2021). Multi-view joint graph representation learning for urban region embedding. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 4431–4437.