


Mapping High Risk Drinking Locations from Different Clustering Methods

João Augusto dos Santos Silva   [Pontifícia Universidade Católica de Minas Gerais | joao.silva.452811@sga.pucminas.br]

Felipe D. da Cunha  [Pontifícia Universidade Católica de Minas Gerais | felipe@pucminas.br]

Silvio Jamil F. Guimarães  [Pontifícia Universidade Católica de Minas Gerais | sjamil@pucminas.br]

 Instituto de Ciências Exatas e Informática, Pontifícia Universidade Católica de Minas Gerais, R. Dom José Gaspar, 500, Belo Horizonte, MG, 30535-901, Brazil.

Received: 04 November 2023 • **Accepted:** 22 October 2024 • **Published:** 21 November 2024

Abstract Over the years, there has been a significant increase in the prevalence of diseases associated with the misuse of alcoholic beverages, resulting in three million annual deaths worldwide. Despite this alarming trend, there is a lack of dedicated applications to support individuals in their recovery from alcohol abuse. In light of this situation, the literature presents machine learning techniques that can be employed to identify and characterize urban areas with a high propensity for alcohol consumption in major cities. This study explores the utilization of Location-Based Social Networks (LBSN) to assess alcohol consumption habits in Tokyo and New York. Data from check-ins at bars and restaurants were collected, and through clustering methods, the study examined the drinking patterns of urban residents. The findings revealed that, while there were cultural variations in drinking behaviors between the two cities, users tended to consume more alcohol during weekends and nighttime. Furthermore, the research successfully pinpointed the regions most conducive to such consumption.

Keywords: Graph, Clustering, LBSN, Smart City, Pervasive Computing

1 Introduction

In recent years, society observed an increasing number of illnesses and deaths caused by abusive consumption of alcoholic beverages, the leading cause of 200 diseases and causing 3 million deaths around the world per year, which corresponds to 5% of global deaths, according to the World Health Organization [WHO, 2022]. From these data, the comprehension of causes and consequences for the elevated level of consumption can be an essential feature to help authorities take necessary measures to combat alcoholism, such as developing new medications for this purpose, as described in Han *et al.* [2021] or studying the consumption patterns as mentioned in Boschuetz *et al.* [2020].

Nowadays, there are special studies for the fight against the high consumption of alcoholic drinks, as described in Dulin *et al.* [2014] and Gustafson *et al.* [2014], that proposed applications capable of performing digital monitoring of people who are recovering from alcohol abuse through functionalities projected to serve this audience. The identification of **High-Risk Drinking Locations** (or **vulnerable regions**) could be considered as the main functionality of those applications. The functionality consists of the register of locations where the user had the habit of drinking alcoholic beverages. After registering these locations, the application was supposed to emit alerts when the user was close to these regions, allowing him to stay in this location temporarily. The mapping from the High-Risk Drinking Locations is static, which means that the user needs to create the register from that location in the application, which brought the main objective of the present work, the dynamic classifica-

tion from regions using check-ins collected from Location Based Social Networks (LBSNs), one of the primary sensors in urban computing.

Urban computing uses many data sources, such as the Internet of Things (IoT) devices, Location-Based Social Networks (LBSN) data, and statistical data, which facilitate understanding the urban environment. As mentioned in Silva *et al.* [2014b], Rodrigues *et al.* [2019] and Skora and Silva [2021], urban computing can make a difference in various areas. With the increasing availability of data through smart city implementation initiatives, individuals are monitored in various aspects, such as mobility, routines, interests, feelings, and more. All these collectible data provide us with insights into different domains. As highlighted in Machado *et al.* [2015] and Gubert *et al.* [2022], it is possible to explore data from various domains through layered sensing and multi-aspect graphs, which enables the analysis of the influence of factors like traffic and weather conditions on people's mobility based on the social classes of a city and the dynamism in points of interest. Building upon the prior research [Silva *et al.*, 2023a], our primary goal in this work is to propose a **strategy capable of delineating urban areas based on their predominant activities** and showing that our strategy works. We create an application to illustrate the results. Moreover, we can say that the main focus here is to identify regions with a high likelihood of consuming alcoholic beverages since this information may serve various purposes, including facilitating data-driven marketing strategies related to alcohol, optimizing delivery route planning, aiding regulatory oversight, and assisting individuals in alcohol recovery programs to avoid triggering environments.

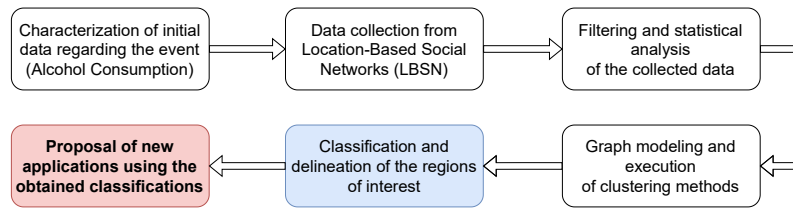


Figure 1. Pipeline for Mapping High-Risk Drinking Locations. Here we illustrate the methodology applied in this work.

Thus, the main contributions of this work are three-fold: (i) we propose a new graph-based strategy for delimiting vulnerable regions, taking into account distances between the points of the regions and the size of the regions; (ii) a categorization of these urban regions relies on data gathered from Location-Based Social Networks, and (iii) a comparison of this proposed method to two others. Regarding categorization, these data, after undergoing the necessary preprocessing, are subjected to three distinct methods to separate them. We will define these methods as clustering for simplicity and without loss of generality. The first method uses graph analysis, while the other two leverage classical machine learning techniques. These clustering techniques yield clusters that are subsequently analyzed and categorized. It provides a classification for the clusters and enhances our understanding of the movement patterns of individuals within the urban landscape.

The work is organized as follows. Section 2 discusses the most relevant works that motivated this study and the chosen methodology. Section 3 delves into the methodology, covering data collection, the databases used, and tools. Section 4 presents the clustering methods used in this work. Section 5 presents the results obtained in region classification. Section 6 explores potential applications of the study findings. Finally, in Section 7, we draw some considerations of this study and suggest new directions for future research.

2 Related Work

In the current global context, with the high availability of data collected from social sensors, several authors have presented solutions related to urban mobility. This section presents some works that utilize this data to understand the urban environment.

Meanwhile, in their respective works Dulin *et al.* [2014] and Gustafson *et al.* [2014], the authors have made significant strides in addressing the challenges that patients with alcohol use disorder face during their treatment journey. They have introduced two applications, “StepAway” and “A-CHESS”, which empower users to monitor their progress following alcoholism treatment. These applications, with their unique functionalities, notably the identification of High-Risk Drinking Locations, have made a substantial impact. However, their reliance on user input for defining these locations may only partially capture the dynamic nature of urban environments. In contrast, the current research dynamically categorizes these High-Risk Drinking Locations using Location-Based Social Networks (LBSNs) data.

In Zhang *et al.* [2021], the authors developed algorithms aiming at a multi-view learning model for embedding ur-

ban areas using graphs obtained from data regarding human mobility within cities and attributes of the analyzed regions, which include Point of Interest (POI) and human mobility within the regions. Starting from the initially generated graphs, the Graph Attention Network (GAT) technique is applied to the model to learn about the representativeness of the vertices. Finally, with the obtained results, a joint learning model was created to enable the collaboration of different visualizations.

Lastly, in Le Falher *et al.* [2021], the authors use the check-ins on social networks to study similarities between neighborhoods to provide users with recommendations on which neighborhoods to visit based on their preferences. For instance, if users want to go shopping, the application can direct them to a specific neighborhood, and if they want to dine out, it will suggest another neighborhood. Drawing inspiration from Le Falher *et al.* [2021], the present work aims to combine the functionality of classifying regions with the functionality presented in Dulin *et al.* [2014] and Gustafson *et al.* [2014] for identifying High-Risk Drinking Locations within cities.

Table 1 shows how this research distinguishes itself from others by conducting analyses utilizing check-in data from Location-Based Social Networks (LBSNs), consolidated into a proprietary database described in Section 3.3. The primary focus is on identifying regions that facilitate alcohol consumption. This analysis opens the door to the potential development of novel applications that seamlessly merge health and urban computing, aligning with the suggestions presented in Dulin *et al.* [2014] and Gustafson *et al.* [2014]. This integration can take place dynamically and independently of the application users themselves.

In this study, as we will see in the next, three clustering algorithms are applied to delineate the regions earmarked for analysis. This technique has also been implemented in Le Falher *et al.* [2021] and Zhang *et al.* [2021]. However, in this specific investigation, the K-Means and DBSCAN algorithms are employed in conjunction with a graph-based algorithm inspired in Cousty and Najman [2011] and Felzenszwalb and Huttenlocher [2004]. These algorithms generate clusters that are subsequently subject to analysis and categorization.

3 High Risk Drinking Locations

The methodology for studying alcohol consumption behavior is illustrated in Figure 1. Data collection is a crucial step in the process. The following subsections will discuss all the stages of the methodology used in this study.

| Related Works | Maps High Risk Drinking Locations | Classify clusters in cities | Comparison of clustering models | Graph-based Clustering Algorithm | LBSN data usage |
|-------------------------|-----------------------------------|-----------------------------|---------------------------------|----------------------------------|-----------------|
| Dulin et al. [2014] | X | | | | |
| Gustafson et al. [2014] | X | | | | |
| Zhang et al. [2021] | | | | X | X |
| Le Falher et al. [2021] | | X | | | X |
| Our Work | X | X | X | X | X |

Table 1. Comparison between related works.

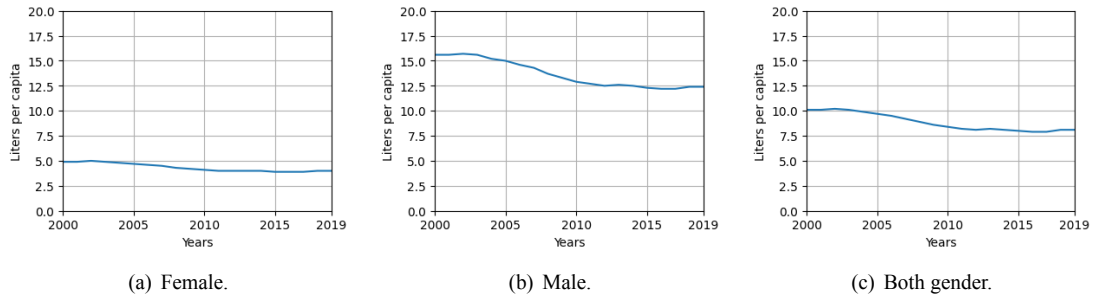


Figure 2. Alcohol consumption in liters per capita (aged 15+) per year in Japan.

3.1 Database

Foursquare¹ is a globally recognized social network that allows users to share their real-time location and feedback about places with their network of friends within the platform or with their followers on linked social networks, such as Twitter² and Facebook³. Over time, Foursquare transitioned the check-in functionality to another dedicated app within the company, Swarm, which offers the same interaction users had within Foursquare but with a platform entirely focused on this feature. According to Foursquare’s privacy policy, check-ins are considered private information. However, some users share their location via Twitter, making check-ins public and allowing access to the data without violating the rules of any of the used social networks, as considered in Silva *et al.* [2014a].

By obtaining check-ins through the Twitter API⁴, approximately 2.7 million instances were collected between May 2022 and January 2023 from around the world, enabling a more comprehensive analysis of each country’s behavior. A more detailed data collection version is in Silva *et al.* [2023b]⁵. In JSON 1, an example of the return obtained through the Twitter API is represented. It includes the Twitter tweet ID, the published text (which contains the URL related to the check-in on the Foursquare Swarm network), the date and time corresponding to the tweet’s registration time by the user, and finally, information regarding the requests made to the Twitter API. From all attributes presented in JSON 1, only the URL on the text field and the timestamp are taken forward to assemble the database described in Table 2. This way, any trace-back can relate a check-in to a tweet since we discard the URL after it is processed.

For the analysis and understanding of alcohol consumption, a database provided by the World Health Organiza-

tion (WHO)⁶ through the Global Health Observatory program was used. This database contains data from the Sustainable Development Goals (SDG)⁷, which encompass 17 objectives defined by United Nations partner nations to eradicate poverty and inequality, improving health, justice, and prosperity, as well as enhancing planet conservation. The Indicator used to understand global alcohol consumption was Indicator 3.5.2, which is linked to SDG 3, aiming to ensure health and well-being for all ages, Target 5, which seeks to strengthen substance abuse prevention and treatment, and Indicator 2, which represents the amount of alcohol consumed over the years by individuals aged 15 and above in various countries.

The graphs illustrated in Figure 2 represent data related to Indicator 3.5.2 for Japan. Unfortunately, forecasting for the coming years is difficult, as no suitable pattern was observed for such predictions. It is important to emphasize that all the graphs for all countries behave similarly. Furthermore, such a forecast would likely exhibit significant variations due to the global context of the COVID-19 pandemic. The chosen dataset within the previously explained Indicator presents the

JSON 1 Response from Twitter API.

```

{
  "data": {
    "id": "1560008545080430600",
    "text": "I'm at Avenida Brasil in Belo Horizonte, MG https://t.co/5FBbhBurSi",
    "timestamp": "2022-08-17 21:00:01.327915"
  },
  "matching_rules": [
    {
      "id": "1559949275353825281",
      "tag": "swarm"
    }
  ]
}

```

¹<https://foursquare.com>

²<https://twitter.com>

³<https://www.facebook.com>

⁴<https://developer.twitter.com>

⁵Database is available at <https://doi.org/10.5281/zenodo.10037884>

⁶<https://who.int/>

⁷<https://sdgs.un.org/>

volume of alcohol consumed in 186 countries between 2000 and 2019. It accounts for individuals aged 15 and above, divided by gender, showing the difference in consumption between men and women and the average for both.

3.2 Tools

This work used various platforms to develop the necessary tools for each stage of studying consumption behavior. Figure 3 shows all platforms used at this work in the respective order applied to the development. The Twitter API was used for collecting check-ins, utilizing the free license that allowed the monthly collection of 500,000 tweets. The data was collected on a virtual machine hosted on Microsoft Azure⁸ using code written in Python. The programming language is also used for processing the collected check-ins, data preprocessing, database assembly, and analysis within the Jupyter Notebook and Google Colab platforms. The clustering algorithms for region characterization were executed on these platforms as well. All the code used in this work is available in a repository at GitHub⁹.

3.3 Preprocessing

Recognizing that the database contains more information than necessary for our analysis, we removed irrelevant and sensitive data from the collected tweets. We developed comprehensive filters for this data treatment to ensure the utmost accuracy. Our Python-based application was instrumental in parsing the HTML code associated with the check-in link in Swarm, allowing us to extract the necessary data from each check-in. This process facilitated retrieving the desired information and eliminated the need to rely on the Places API from the Foursquare platform. As a result of this thorough data filtering process, we were able to create an initial database, as detailed in Table 2, containing approximately 2.7 million instances. However, after removing rows with missing data, the filtered database now contains approximately 1 million instances, as shown in Table 3, which displays the number of check-ins in the studied cities and the number of venues in each city.

Table 2 shows all attributes inserted on our database, with attributes referring to the specific venue, such as `venueID`, `venueName`, `category`, `country`, `city`, `latitude`, and `longitude`. We also have an ID referring to the user, but we can not trace back the ID for the owner from the check-in, and we also have our temporal attribute called `timestamp`.

4 Clustering

To generate clusters to classify cities based on regions, three data clustering algorithms were used: (i) a density-based algorithm (DBSCAN), (ii) K-Means, and (iii) a graph-based clustering algorithm.

| Attribute | Attribute Description |
|------------------------|--------------------------------|
| <code>venueID</code> | Venue identifier on Foursquare |
| <code>userID</code> | User identifier on Swarm |
| <code>venueName</code> | Venue's name |
| <code>category</code> | Venue's category |
| <code>country</code> | Venue's country |
| <code>city</code> | Venue's city |
| <code>timestamp</code> | Timestamp from post on Twitter |
| <code>latitude</code> | Venue's latitude |
| <code>longitude</code> | Venue's longitude |

Table 2. Database description.

| City | Check-ins | Venues |
|----------|-----------|---------|
| Tokyo | 17,320 | 2,635 |
| New York | 1,916 | 1,001 |
| Others | 966,104 | 230,511 |
| Total | 985,661 | 234,147 |

Table 3. Check-ins collected per city after processing.

4.1 Density-based algorithm – DBSCAN

DBSCAN, which stands for Density-Based Spatial Clustering of Applications with Noise, is a machine learning algorithm focused on density-based clustering. We chose this algorithm because it works with dense data, which we observed in our database. It evaluates the distance between points from a random starting point to differentiate between the found groups, allowing the algorithm to determine the number of generated clusters after its execution. The algorithm proposed in Ester *et al.* [1996] has drawbacks, as depending on the choice of the starting point, it may generate different clusters. Additionally, DBSCAN does not work well with large datasets, making scalability impractical. On the other hand, it delivers excellent results by classifying outliers as noise, eliminating the need for their removal during data pre and post-processing.

4.2 K-Means

K-Means is an unsupervised machine learning algorithm designed for data clustering based on their characteristics MacQueen *et al.* [1967]. Unlike the DBSCAN algorithm, applying K-Means requires the prior definition of the desired number of clusters. We used the Elbow method to define this number, as represented in Figure 4. The Elbow method tests the data's variance concerning the number of clusters to return the ideal number for the clustering. In Figure 4(b), you can identify the inflection point on the graph near the value 3 on the x-axis, indicating that beyond this value, there will be no gain in clustering with an increasing number of clusters. The same analysis can be performed in Figure 4(a), representing the Elbow for New York City, where the inflection point approaches the value 2 on the x-axis. The K-means algorithm was chosen because of its simplicity and for its dissemination of the clustering community.

4.3 Graph-Based Clustering Algorithm

A graph $G = (V, E)$ consists of a finite set of vertices, denoted by V , and a finite set of edges denoted by E , in which $E \subseteq V \times V$. If $\{u, v\} \in E$ for two vertices $u, v \in V$, then u

⁸<https://azure.microsoft.com>

⁹<https://github.com/joaogustoss/High-Drinking-Sense>



Figure 3. Flowchart of the used tools. Here, we present the tools we used step-by-step to develop this work.

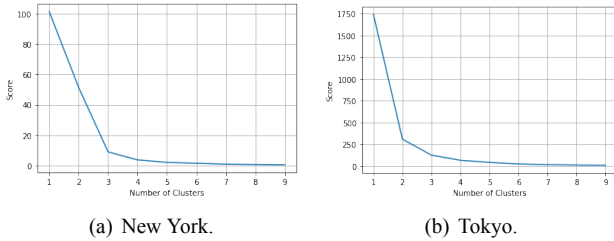


Figure 4. Elbow method indicating ideal number of clusters for New York and Tokyo.

and v are adjacent vertices. The notion of vertices relates to the data's elemental components and edges and the connections and dynamics between the parts.

Without loss of generality, the graph partitioning could be a clustering method in which each partition region may be a cluster. A partition \mathbb{P} is a set of non-empty disjoint subsets of V , meaning that $\forall X, Y \in \mathbb{P}$, X and Y are regions, $X \cap Y = \emptyset$ if $X \neq Y$ and $\cup\{X \in \mathbb{P} = V\}$.

In the case of modeling the venue clustering, a venue will be a vertex of the graph, and two venues are connected if it is possible to reach one another. Moreover, the edge weights are the distance between two connected venues calculated from their geographical coordinates. Instead of getting a complete graph, we have used a KD-Tree (K-Dimensional Tree) [Ooi, 1987] data structure using Euclidean Distances to identify the k -nearest neighbors. To define the k parameter indicating the number of connections, we considered the need to obtain a connected graph with only one component. In New York, we used $k = 6$, and in Tokyo, $k = 11$.

In Felzenszwalb and Huttenlocher [2004], the authors presented a graph-based algorithm for image segmentation in local neighborhoods. Felzenszwalb's Algorithm aims to partition an image into meaningful regions or segments based on color similarity and proximity of pixels. In its operation, the algorithm needs an input image to be segmented and a few parameters: the sigma or scale, which controls the size of resulting regions; the threshold, which also controls the size and the quality of segmented regions; and the last parameter called min size, that controls the minimum number of pixels desired in a region.

Following the abovementioned definitions, we developed an algorithm based on the Watershed Hierarchy Cuts for image segmentation [Cousty and Najman, 2011]. We aim to generate clusters with the minimum vertex difference in the developed algorithm. For this, we iterate through all the edges on the graph and calculate how many vertex will be contained in each component after the edge's removal if the difference between the vertex set cardinality of the two connected components is lower than the previous value obtained by the removal of another vertex, this value is updated, and we save the edge's ID from removing it after verifying all the edges on the graph. After removing an edge and obtaining two connected components, we repeat the process in each connected component to obtain more clusters. The computational cost of our algorithm is $O(N)$, where N corresponds

to the number of vertex on the graph.

4.4 Comparing Clustering Algorithms

The DBSCAN algorithm was applied to the two selected cities for analysis, considering only the geographical coordinates of the data. The results were unsuitable, even after adjusting the 'eps' hyperparameter, which defines the required distance between two points considered in the same group. For clustering the data related to the city of Tokyo, as illustrated in Figure 5(d), it was necessary to set the 'eps' to a value of 0.02, resulting in clustering with 4 clusters and 1 group containing 8 instances classified as noise. In the case of New York City, an 'eps' value of 0.02 was also used, generating clustering with 2 clusters and 1 group containing 50 instances classified as noise, as illustrated in Figure 5(a). The groups classified as noise are in cluster -1 in Figures 5(a) and 5(d).

After identifying the geographical coordinates needed to run the K-Means algorithm, we obtained results that were consistent with DBSCAN's results. The main difference lay in classifying clusters with greater check-ins, primarily in city centers, as depicted in Figure 5. It demonstrates the difficulty presented by DBSCAN in clustering very dense data. To define the number of clusters, $k=4$ was chosen for clustering instances related to Tokyo, as shown in Figure 5(e). The result visually differed from DBSCAN, which clustered the data into 4 clusters. In the execution with data related to New York City, represented in Figure 5(b), the value of $k=4$ was also used.

The graph-based clustering procedure is based on sequential removal of the edges, so after removing three edges from the MST , we obtained four connected components, also called clusters. We defined the number of 4 clusters to compare the results with other clustering algorithms previously mentioned in this work. The obtained clustering in New York, Figure 5(c), shows how our algorithm separated the densest part of the city, standing with a similar number of venues in each cluster, similarity K-means did not assure. Moving to the resulting cluster in Tokyo, as seen in Figure 5(f), we can observe a similar behavior to what occurred in New York, where the algorithm could split the densest part of the city into different clusters and kept the similarity between the number of venues in each one.

Table 4 presents some comparisons in terms of venues in each New York and Tokyo cluster. Moreover, the Silhouette Coefficient [Rousseeuw, 1987] computed for each algorithm and each city is also illustrated. This comparison shows how the developed graph-based clustering algorithm could minimize the difference between clusters by choosing the ideal edge to be removed. However, it demonstrated a low score for the Silhouette Coefficient compared to both executed algorithms. As a result of this selection, we obtain more balanced clusters, which can impact the final classification result and lower the evaluated metric for cluster similar-

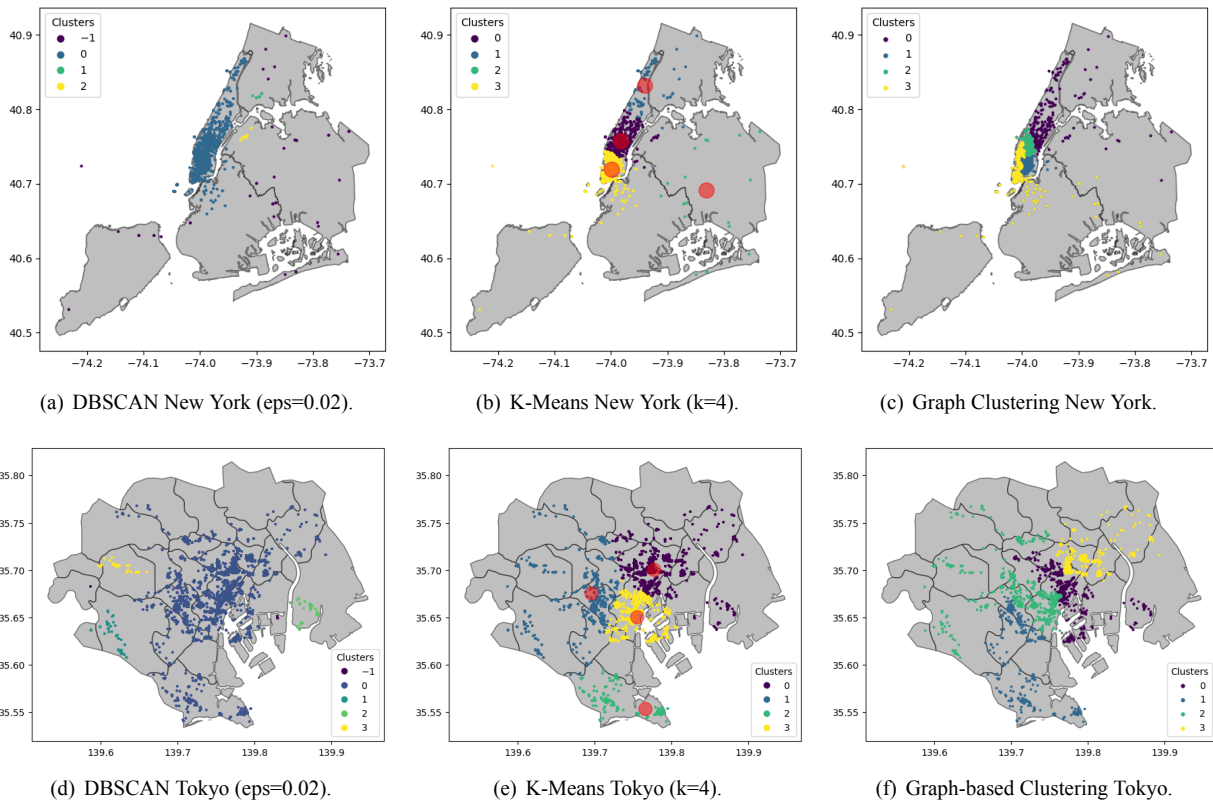


Figure 5. Clustering methods results. Here, we illustrate the results computed by DBSCAN, K-Means, and Graph-based methods, considering data from New York and Tokyo.

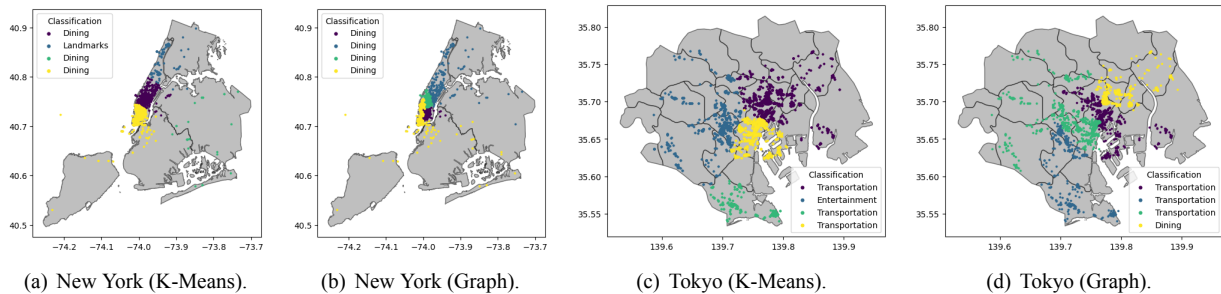


Figure 6. Georeferenced classification results. Here, we illustrate the results computed using the graph-based method, considering data from New York and Tokyo.

ity; since our algorithm does not aim to obtain similar clusters based on geographical coordinates, our goal is to obtain similar clusters in the number of venues.

In summary, while DBSCAN presents severe difficulty in clustering very dense data, the graph-based method has successfully separated the densest part thanks to its ability to identify reasonable distances between clusters. When comparing the K-Means to the graph-based method, despite the better scores for the silhouette coefficient obtained by the K-Means against the graph-based method, the proposed method produced more balanced clusters regarding the number of venues.

5 Experimental results

In this section, the results of the neighborhood classifications obtained from the clusters generated by the Graph-based clus-

tering algorithm and K-Means, as exemplified in Section 4, are evaluated and compared. Our classification process involved several steps. We began by considering the category attribute from the database. We then counted the most frequent category in each generated cluster and considered it the classification for that specific cluster. To provide a clearer picture, out of the initial 18.7 thousand data points, we obtained 1.9 thousand check-ins in 1.1 thousand unique venues in New York, divided into 4 regions. In Tokyo, we obtained 17.3 thousand check-ins in 2.6 thousand unique venues, also divided into 4 regions.

The analysis of the data obtained after clustering was divided into two parts: spatial analysis, where we considered only georeferenced data related to check-in publication for cluster classification (Section 5.1), and temporal analysis, which considered information related to the day and time of check-in publication, dividing the data into weekdays and periods (Section 5.2). This division of analyses allows us to un-

| Cluster | Graph-based | K-means | DBSCAN |
|------------------------|-------------|---------|--------|
| -1 | 0 | 0 | 31 |
| 0 | 261 | 479 | 959 |
| 1 | 277 | 66 | 5 |
| 2 | 226 | 18 | 6 |
| 3 | 237 | 438 | 0 |
| Silhouette Coefficient | 0,40 | 0,50 | 0,15 |

(a) Number of venues per cluster in New York

| Cluster | Graph-based | K-means | DBSCAN |
|------------------------|-------------|---------|--------|
| -1 | 0 | 0 | 4 |
| 0 | 553 | 1131 | 2518 |
| 1 | 563 | 771 | 43 |
| 2 | 754 | 185 | 29 |
| 3 | 765 | 548 | 41 |
| Silhouette Coefficient | 0,28 | 0,48 | 0,20 |

(b) Number of venues per cluster in Tokyo

Table 4. Number of venues per cluster in Tokyo and New York.

derstand human mobility in the analyzed regions, making it possible to comprehend differences in human behavior based on the day of the week and time of day. The possible classifications include Entertainment, Business, Community, Dining, Event, Health, Landmarks, Retail, Sports, Transportation, or Residence, which correspond to the categories found on Foursquare related to each venue.

5.1 Georeferenced Analysis

Following the delineation of regions produced by Graph-based clustering and K-Means, the resulting regions underwent a classification process. In this initial analysis, we calculated the count of check-ins at each location within a specific cluster. Once we had this count, we examined the category with the highest number of check-ins, and based on this category, the cluster was categorized.

In Figure 6, the georeferenced classification obtained in Tokyo and New York is compared using the clustering algorithms studied. The classification is consistent between the two cities, with few discrepancies observed in the cluster 1 in New York and the clusters 1 and 3 in Tokyo. However, this initial analysis does not accurately capture the dynamics of cities because it is a static assessment that does not consider a crucial variable in urban dynamics – time. In the following section, we incorporated time information into two distinct approaches to improve the classification.

5.2 Temporal Analysis

From the georeferenced analysis, we can affirm that the day and time at which the user checks in are highly relevant in classifying the region sought by users, which can define momentary points of interest, such as events in cities, including professional events that usually occur during the week

and business hours or leisure events, such as shows that typically take place on weekends and during the evening. The detection of these points of interest is carried out by the high concentration of check-ins within a short period at the same venue or in geographical coordinates close to each other, as mentioned in Silva *et al.* [2013]. The analysis considering temporal data related to the day and time of check-in is divided into two parts: the analysis considering the day of the week in Section 5.2.1 and the analysis with data related to the time in Section 5.2.2.

5.2.1 Weekdays

The temporal information was divided into groups to analyze and classify regions based on the day’s users checked in. The first group corresponds to weekday check-ins (from Monday to Friday), and the second corresponds to weekends (Saturday and Sunday).

From the data illustrated in Figure 7, it was possible to visualize differences between the classifications performed in New York and Tokyo. Figures 7(a), 7(b), 7(e) and 7(f) shows the same classification comparing the regions obtained from K-Means and our Graph-Based clustering method. This result shows that both algorithms could show the differences between the population’s interests on weekdays and weekends. The same can be observed in Figures 7(c), 7(d), 7(g) and 7(h), related to the classifications performed in Tokyo. The region classifications obtained from our clustering method differed from K-Means, where all regions were Transportation on weekdays.

Interestingly, our clustering method classified the region on the city’s East side as Dining and Drinking, a classification that was not expected. Regarding the weekend classification, both clustering methods obtained the same classification for the East side of Tokyo. However, they differed on the West side, with the Entertainment classification on the K-Means region and Transportation on our clustering method.

This analysis shows different results than the georeferential analysis in Section 5.1. However, it is still incapable of showing the intense dynamism of the urban environment, which brought the idea of analyzing classifications based on the hour of the day.

5.2.2 Time

In addition to identifying points of interest, the temporal analysis can also provide insights into the lifestyle and behavior of cities. Figure 8 displays the classifications of regions in New York City for both clustering methods, and it reveals a distinct dynamic compared to the classifications discussed in Section 5.2.1. This contrast is particularly evident in the period from 12 : 00 to 00 : 00, during which all regions formed by both clustering methods are uniformly categorized as “Dining and Drinking”, which indicates that after 12 : 00, the predominant venue category in all regions is Dining and Drinking.

When examining the classifications obtained in Tokyo, it becomes apparent that the time-based classification can better represent the city’s dynamics. In the period from 00:00 to 12 : 00, the predominant category in Tokyo is “Travel

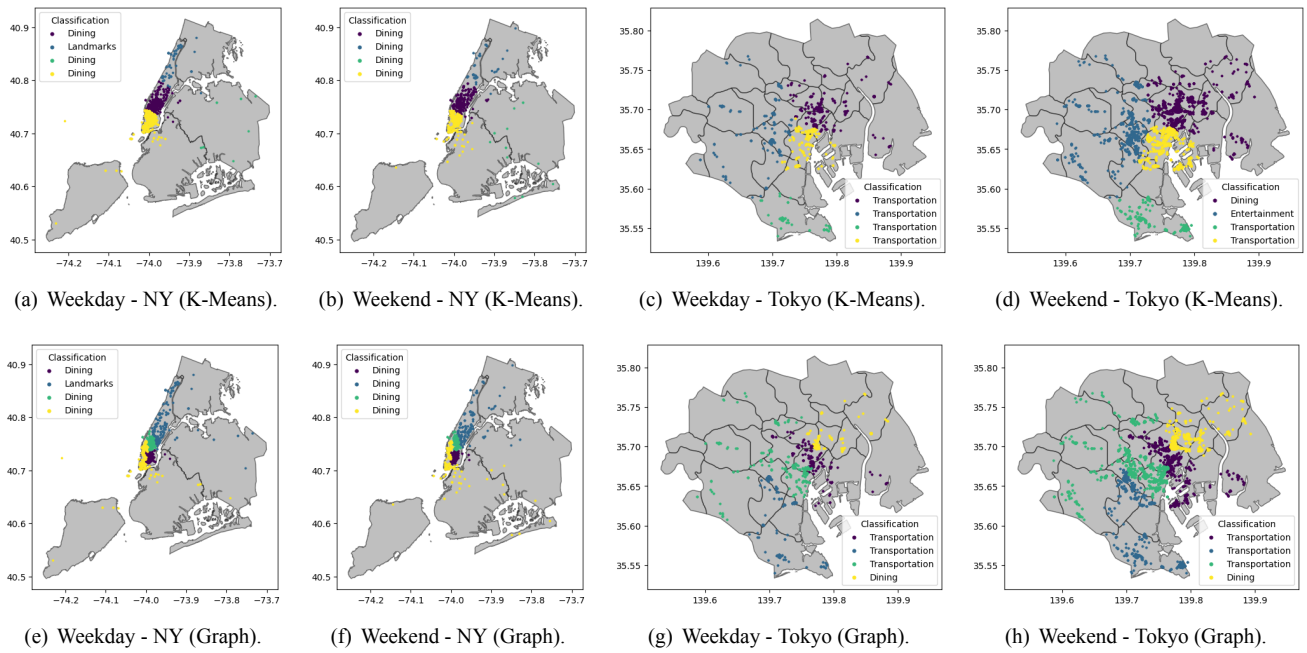


Figure 7. Region classification results considering the day of the week. Here, we illustrate the results computed using the K-Means and graph-based method, taking into account data from New York and Tokyo.

and Transportation”, which aligns with the truth, as this is the time when people typically commute from home to work. The remainder of the day in Tokyo exhibits a dynamic pattern of classifications. For example, between 12 : 00 and 18 : 00, the categories “Arts and Entertainment” and “Dining and Drinking” emerge in regions previously classified as “Travel and Transportation”. During the final quarter of the day, from 18 : 00 to 00 : 00, the “Travel and Transportation” classification is present in three regions delimited by our clustering method, while it appears twice in K-Means regions.

After investigating all the macro-categories mentioned above, we decided to dive into the subcategories within the broader classification of “Dining and Drinking” to pinpoint the High-Risk Drinking Locations in the cities. Figures 10 and 11 illustrate the classifications of subcategories within “Dining and Drinking” in New York and Tokyo, respectively. In New York, we observe significant variations in the classifications. During the first half of the day, all regions are categorized as “Bar” in our clustering method. In contrast, K-Means-generated regions include one region with no check-ins recorded from 00 : 00 to 06 : 00. Furthermore, K-Means designates some regions as “Bar” during this time range, which implies that these regions during this time frame could be considered High-Risk Drinking Locations, much like all regions in our clustering method.

Analyzing the second time range, from 06 : 00 to 12 : 00, Figures 10(f) and 10(b), the classification Cafe is pronounced, classifying three regions from K-Means clusters and two regions from our clustering, what makes sense, since it is the time when people are looking for breakfast before your work hours. The other classifications in this time range are Restaurant and Smoothie Shop. The category Restaurant is predominant when analyzing region classification from

12 : 00 to 00 : 00, except for one region from K-Means, which was classified as Bar from 12 : 00 until 18 : 00.

The classifications incorporating check-in times have proven more effective than those detailed in Sections 5.1 and 5.2.1. This enhanced effectiveness is attributed to the discernible variations in the classifications, which align with the continuous movement of people and their evolving preferences according to the time of day. These findings are particularly pertinent to cities characterized by a high inrush of people throughout the day.

6 Applications

Based on the classifications presented in Section 5 and with the clustering performed in Section 4, it is possible to propose applications in the following areas:

Enforcement: Guiding public agencies responsible for traffic enforcement and overseeing the regulation of substances not permissible for minors to specific regions categorized under “Dining and Drinking”. This strategic approach aims to proactively deter potential criminal activities and promote a safer environment within these areas.

E-Health: Detecting areas where alcohol consumption is prevalent, thereby highlighting potential hazards for individuals in the process of recovering from alcohol addiction, with the overarching goal of preventing relapses and providing necessary support.

Logistics: Facilitating logistics and supply chain planning to promptly deliver goods to food and beverage establishments, ensuring efficient operations and uninterrupted services.

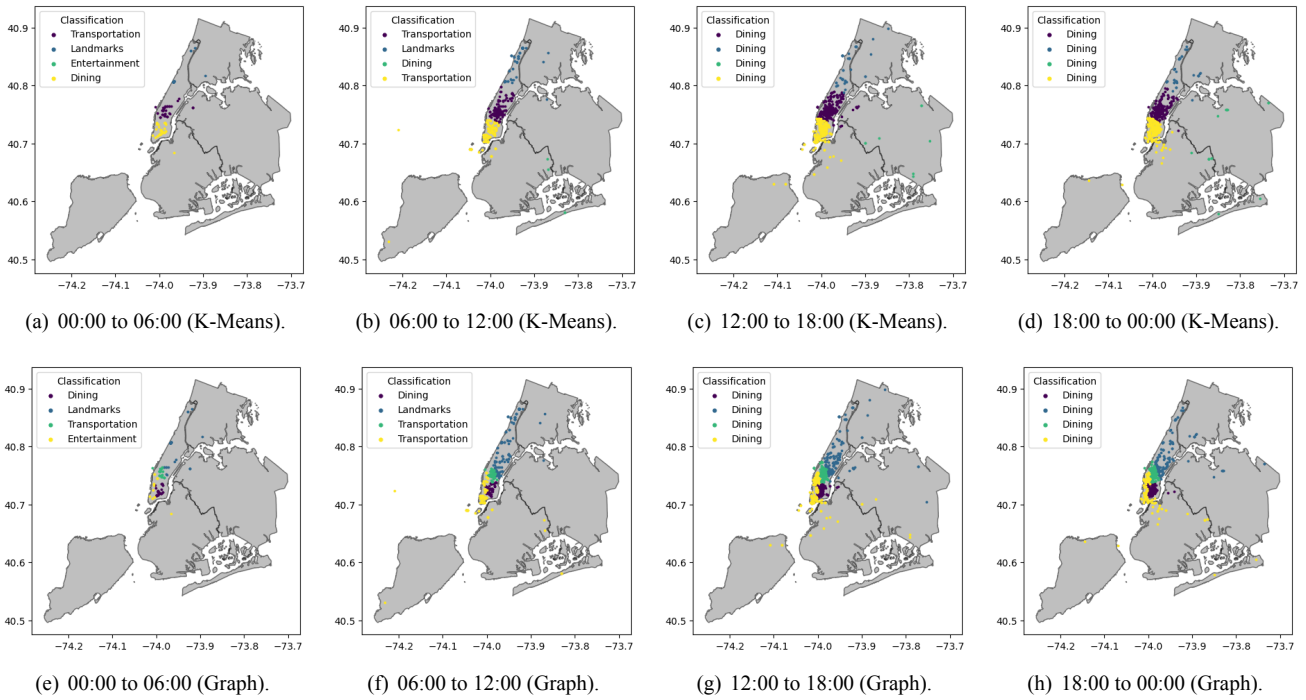


Figure 8. Regions classification results based on the time of check-in. Here, we illustrate the results computed using the K-Means and graph-based method, considering data from New York.

Marketing: Creating a geographical map of areas where companies and businesses can optimize their marketing efforts to achieve heightened success and reach their target audience more effectively.

Traffic Management: Strategically planning vehicular routes to circumvent areas with a dense concentration of check-ins, which may serve as early indicators of potential traffic congestion.

7 Conclusion

Considering the findings detailed in Section 5 on alcohol consumption, it is necessary to explore methodologies that can offer support to people seeking to abstain from alcoholic beverages. Additionally, there is a need for tools to assist in monitoring, marketing, and logistics related to areas with a high concentration of establishments such as bars, restaurants, and other venues selling alcoholic products. To meet these requirements, this study analyzed check-in data collected from Location-Based Social Networks (LBSNs) using clustering algorithms to identify clusters based on their geographic locations. These clusters were subsequently categorized based on the predominant venue type, with consideration for time of day, day of the week, and an analysis that does not take temporal information into account.

The proposed Graph-Based clustering algorithm is a significant advancement in our research, as it enables the grouping of data points and the subsequent categorization of these clusters according to the dominant venue types within them. In literature, we have Le Falher *et al.* [2021] that also categorizes clusters using different clustering methods. This method's outcomes were quite promising compared to the

traditional K-Means clustering algorithm. This approach allowed for the dynamic identification of “High-Risk Drinking Locations” by pinpointing clusters with a concentration of check-ins during evening hours and on weekends, aligning with the initial objective of this study by adapting what was proposed in Dulin *et al.* [2014] and Gustafson *et al.* [2014], by mapping the “High-Risk Drinking Locations” dynamically using LBSN’s. Furthermore, this approach sheds light on the urban mobility patterns within the cities analyzed and provides valuable information on the patterns of alcohol consumption on specific days and times. Despite the marked cultural differences between New York and Tokyo, this study demonstrates the effectiveness of social sensors in depicting urban dynamics and mapping alcohol consumption based on selected days and times.

Looking ahead to future research endeavors, diving into various methodologies for region clustering is imperative. Furthermore, there is substantial merit in extending the assessment of the classifications derived from New York and Tokyo to other urban centers. These evaluations can provide insights into the applicability and adaptability of classifications in various geographical contexts. Additionally, it is vital to consider implementing one of the proposed applications outlined in Section 6, using the regional classifications as a data source. This study expansion can offer valuable insights and applications in many areas of urban environments, such as urban planning, public health, and data analysis.

Funding

This research was funded by Fundação de Amparo a Pesquisa de Minas Gerais (FAPEMIG) (Grant APQ-01079-23, PPM-00006-18 e PIBIC 2022/28009), the Conselho Nacional de Desenvolvi-

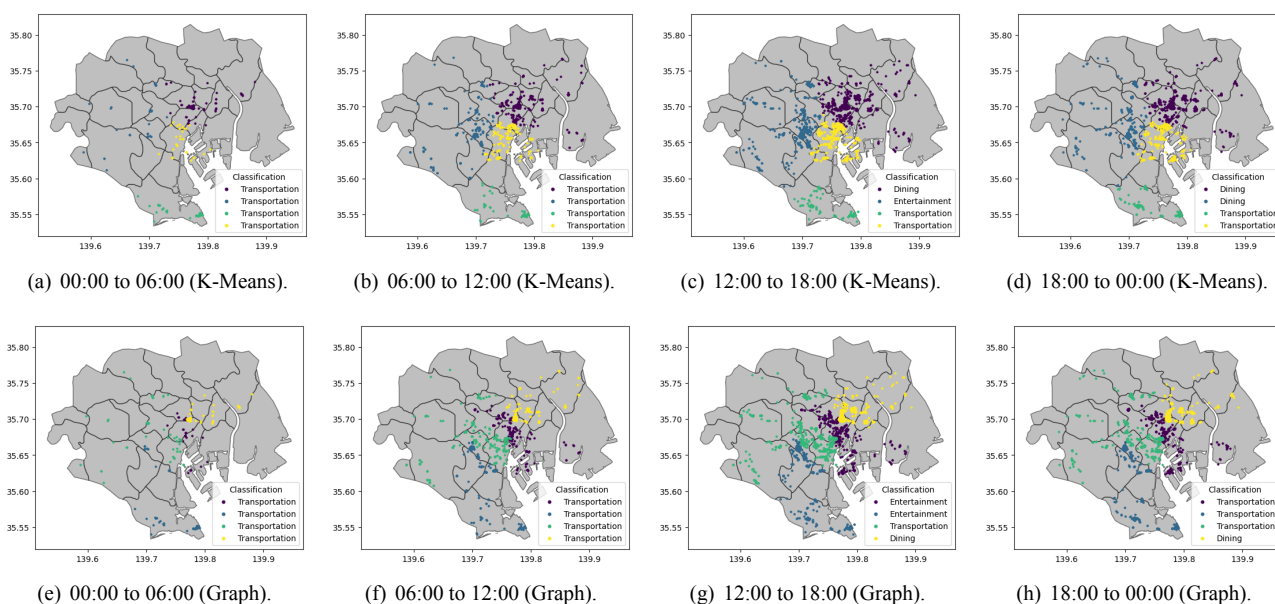


Figure 9. Regions classification results based on the time of check-in. Here, we illustrate the results computed using the K-Means and graph-based method, considering data from Tokyo.

mento Científico e Tecnológico – CNPq (Grants 407242/2021-0 and 306573/2022-9) and Pontifícia Universidade Católica de Minas Gerais (PUC Minas).

Authors' Contributions

JS contributed to the conception of this study and performed the experiments. FC and SG contributed to the rich discussions and the conception of this study. JS is the main contributor and writer of this manuscript. All authors read and approved the final manuscript.

Availability of data and materials

The datasets generated and analyzed during the current study are available at <https://doi.org/10.5281/zenodo.10037884>. The codes used during the current study are available at <https://github.com/joaogaugustoss/High-Drinking-Sense>.

References

- Boschuetz, N., Cheng, S., Mei, L., and Loy, V. M. (2020). Changes in alcohol use patterns in the united states during covid-19 pandemic. *Wmj*, 119(3):171–176. Available at <https://pubmed.ncbi.nlm.nih.gov/33091284/>.
- Cousty, J. and Najman, L. (2011). Incremental algorithm for hierarchical minimum spanning forests and saliency of watershed cuts. In *International Symposium on Mathematical Morphology and Its Applications to Signal and Image Processing*, pages 272–283. Springer. DOI: 10.1007/978-3-642-21569-8_24.
- 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. DOI: 10.1080/08897077.2013.821437.
- 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. DOI: 10.1023/A:1009745219419.
- Felzenszwalb, P. F. and Huttenlocher, D. P. (2004). Efficient graph-based image segmentation. *International journal of computer vision*, 59:167–181. DOI: 10.1023/B:VISI.0000022288.19776.77.
- 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. DOI: 10.5753/webmedia_estendido.2022.227043.
- 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. DOI: 10.1001/jamapsychiatry.2013.4642.
- Han, B., Jones, C. M., Einstein, E. B., Powell, P. A., and Compton, W. M. (2021). Use of Medications for Alcohol Use Disorder in the US: Results From the 2019 National Survey on Drug Use and Health. *JAMA Psychiatry*, 78(8):922–924. DOI: 10.1001/jamapsychiatry.2021.1271.
- 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. DOI: 10.1609/icwsm.v9i1.14602.
- 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. DOI: 10.1109/MobServ.2015.50.
- MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings*

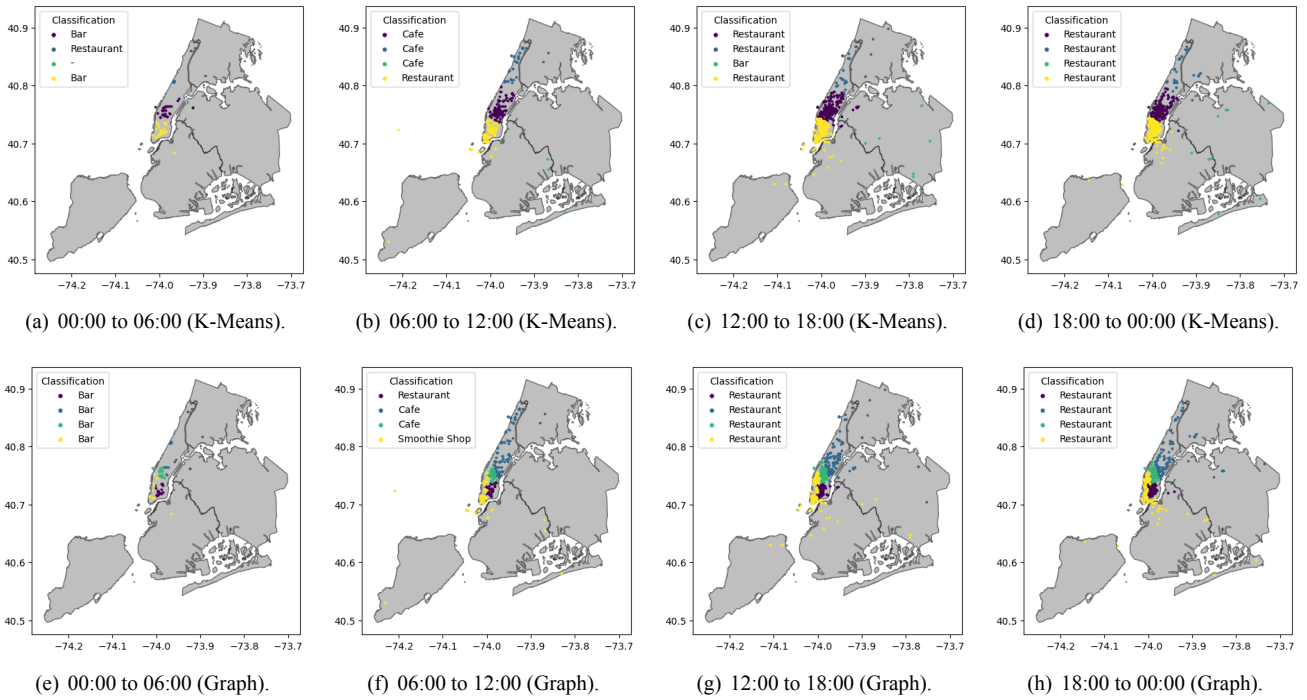


Figure 10. Regions classification results based on the Dining and Drinking category check-in time. Here, we illustrate the results computed using the K-Means and graph-based method, considering data from New York.

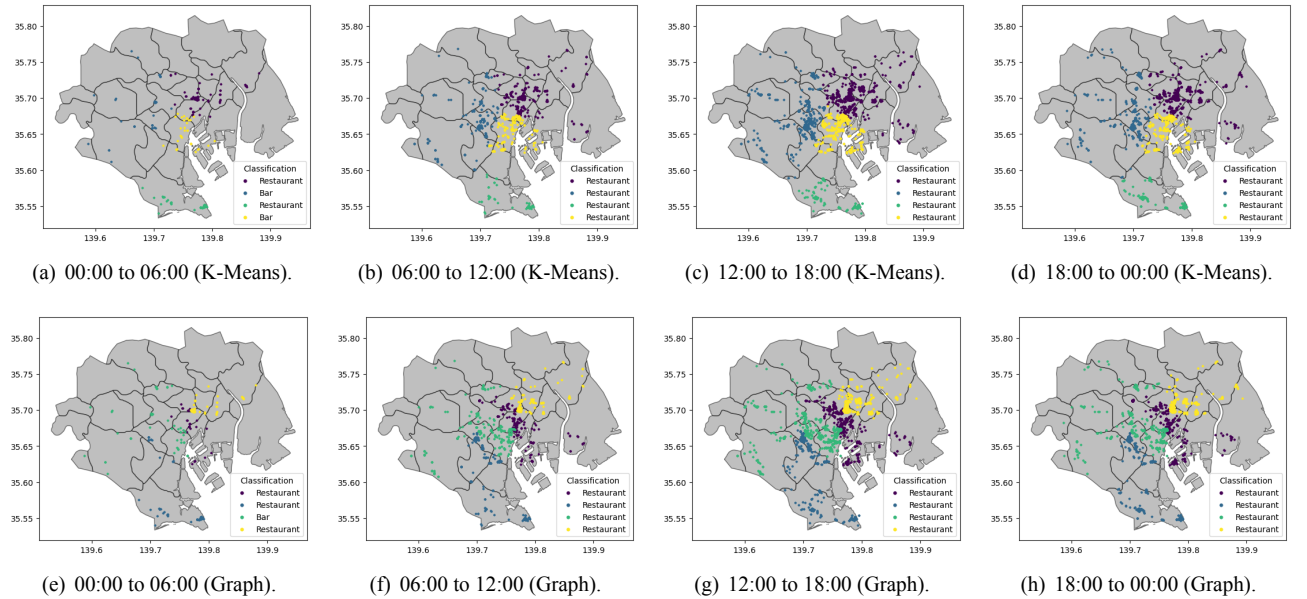


Figure 11. Regions classification results based on the Dining and Drinking category check-in time. Here, we illustrate the results computed using the K-Means and graph-based method, considering data from Tokyo.

of the fifth Berkeley symposium on mathematical statistics and probability, volume 1, pages 281–297. Oakland, CA, USA. Available at <https://projecteuclid.org/ebooks/berkeley-symposium-on-mathematical-statistics-and-probability/Proceedings-of-the-Fifth-Berkeley-Symposium-on-Mathematical-Statistics-and/chapter/Some-methods-for-classification-and-analysis-of-multivariate-observations/bmsp/1200512992>. Ooi, B. C. (1987). Spatial kd-tree: A data structure for ge-

ographic database. In *Datenbanksysteme in Büro, Technik und Wissenschaft: GI-Fachtagung Darmstadt, 1.–3. April 1987 Proceedings*, pages 247–258. Springer. DOI: 10.1007/978-3-642-72617-0_17. 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*. DOI: 10.5753/sbc.6555.9.2.

- Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65. DOI: [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7).
- Silva, J., Cunha, F., and Guimarães, S. (2023a). Estudo do comportamento de consumo de bebida em centros urbanos usando redes de sensoriamento participativo. In *Anais do VII Workshop de Computação Urbana*, pages 68–81, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/courb.2023.774.
- Silva, J. A. S., Cunha, F. D., and Guimarães, S. F. (2023b). Análise da mobilidade urbana por meio de redes sociais baseadas em localização: Estudo de caso em cidades inteligentes. In *Anais Estendidos do XXXVIII Simpósio Brasileiro de Bancos de Dados*, pages 43–49, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/sbbd_estendido.2023.233144.
- 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. DOI: 10.48550/arXiv.1404.1009.
- 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. DOI: 10.1109/DCOSS.2013.59.
- Silva, T. H., Vaz de Melo, P. O. S., Almeida, J. M., Salles, J., and Loureiro, A. A. F. (2014b). Revealing the city that we cannot see. *ACM Trans. Internet Technol.*, 14(4). DOI: 10.1145/2677208.
- 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. DOI: 10.5753/webmedia_estendido.2021.17610.
- WHO (2022). World health organization. Available at: <https://www.who.int/news-room/fact-sheets/detail/alcohol>. Accessed at 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. DOI: 10.24963/ijcai.2020/611.