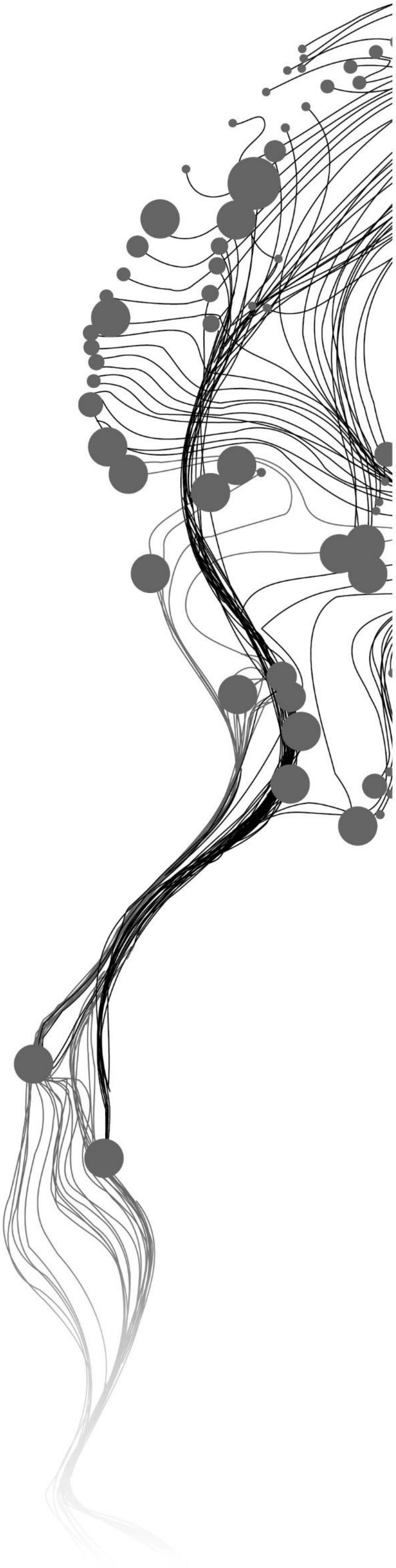
A GRAPH-BASED CLUSTERING APPROACH FOR DETECTING RECURRING SPATIO-TEMPORAL GROUPINGS IN EVENT DATA

RAJIT KUMAR BHAT
September 2021

SUPERVISORS:
Mr. Ashutosh Kumar Jha
Dr. Mahdi Farnaghi



RAJIT KUMAR BHAT

Enschede, The Netherlands, September 2021

Thesis submitted to the Faculty of Geo-Information Science and Earth Observation of the University of Twente in partial fulfilment of the requirements for the degree of Master of Science in Geo-information Science and Earth Observation.
Specialization: Geoinformation science and Earth Observation (M-GEO)

SUPERVISORS:
Mr. Ashutosh Kumar Jha
Dr. Mahdi Farnaghi

THESIS ASSESSMENT BOARD:
Dr. Raul Zurita Milla (Chair)
Dr. Rahul Dev Garg (External Examiner, Indian Institute of Technology, Roorkee

DISCLAIMER

This document describes work undertaken as part of a programme of study at the Faculty of Geo-Information Science and Earth Observation of the University of Twente. All views and opinions expressed therein remain the sole responsibility of the author, and do not necessarily represent those of the Faculty.

ABSTRACT

With a proliferation of location-based devices, a large quantum of data is being generated, the nature of which is spatial, temporal or Spatio-temporal (ST). ST data has benefits over purely spatial or temporal data as it simultaneously helps in understanding persistence patterns and highlight unusual patterns over time. It is observed in mobility studies for humans and animals that movements are not random and serve a certain purpose which is fundamental to the dynamics of the respective ecosystems. Thus, understanding the pattern of their recurring nature can give insights into their groupings in both space and time.

Clustering can be helpful in interpreting ST data. However, clustering algorithms that use Euclidean or planar distance do not account for constraints offered by space and overestimate clustering. Moreover, existing exploratory methods that have a linear view of time, may not recognise the recurring nature of ST events. Also, the spatial boundary and temporal boundary need to be combined in a meaningful manner.

The current work develops an unsupervised exploratory method for detecting recurring Spatio-temporal groupings in ST event data, using a graph-based strategy to agglomerative hierarchical clustering framework. The method is applied to "check-in" data for a location-based social network (LBSN) service to identify clusters or hotspots of users' activity in a city. Hierarchical clustering does not require pre-selection of clusters, enabling automatic discovery of users based on the Spatio-temporal similarity in check-ins. To incorporate the notion of recurrence, a cyclic view of time has been adopted into the temporal distance. The use of a distance metric for clustering where the spatial and temporal components are merged in a weighted linear combination with appropriate scales allows to conceive the spatial and temporal dimensions as required.

Keywords: Spatio-temporal clustering, Hierarchical clustering, Recurrent grouping,

ACKNOWLEDGEMENTS

I wish to take this opportunity to thank the two most important people for their contributions throughout the duration of this research, my supervisors. I am grateful to Mr. AK Jha for his patient supervision and motivation and Dr. M Farnaghi for his consistent encouragement and valuable feedback. While the dream of achieving something very extraordinary was short-lived, I am thankful to my supervisors for pulling me through the days of uncertainty. I appreciate them for the mentorship and help they have extended at every step. Their guidance has helped me grow immensely as a researcher and I will treasure all the advice proffered. Thanks to all the staff at Indian Institute of Remote Sensing and Faculty of Geo-Information Science and Earth Observation, ITC- University of Twente for they played a considerable part during this tenure. Special thanks to Dr. Sameer Saran (HOD-GID, IIRS) for his concern and encouragement. I would also like to thank Dr. Sanders for providing all kinds of support during the coursework. In addition, I wish to thank my classmates at IIRS and ITC for sticking together through both good and trying times. Last but not the least I would like to thank my parents and my sister, for supporting me throughout my thesis work.

Like Calvin once said, "I must obey the inscrutable exhortations of my soul" (*Calvin and Hobbes, by Bill Watterson 1995*). I thank life in general for motivating to push the boundaries and to explore the unknown.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1. Background

Advancement in spatial technologies has led to a proliferation of location-based devices. As a result, a large quantum of data is being generated, which is spatial, temporal or Spatio-temporal (ST)(Oswaldo & Romero, 2011). ST data has both spatial and temporal information. This has benefits over purely spatial or temporal data since activities in real life are carried out over space and time. Spatio-temporal data simultaneously helps in understanding persistence patterns in a phenomenon along with highlighting unusual patterns over time. Hence extracting and interpreting information from ST data is of interest due to widespread applications across different domains.

Researchers have contributed to the development of techniques for discovering potentially meaningful information from ST data to support decision making, also known as geographic knowledge discovery (GKD) (Shekhar et al., 2015). Based on the output obtained from such techniques, a set of algorithms is concerned with detecting patterns or processes based on grouping similar objects by partitioning the underlying space and time, also known as Spatio-temporal clustering (Shekhar et al., 2015).

This study contributes to developing an unsupervised exploratory clustering method to detect recurring Spatio-temporal groupings in ST event data. A graph-based strategy has been adopted to agglomerative hierarchical clustering framework for identifying recurring Spatio-temporal groupings incorporating appropriate spatial and temporal scales. The method is applied to 'check-in' data for a location-based social network (LBSN) service to identify clusters or hotspots of users' activity in a city.

1.2. Motivation-

According to the *Encyclopaedia of GIS* (2017), "Movement patterns associated with the activity of objects can be viewed as a Spatio-temporal expression of their behaviour"(Gudmundsson et al., 2008). Studies in mobility for humans and animals acknowledge movement through space over time to perform activities. It was observed that these movements are not random and serve a certain purpose which is fundamental to the dynamics of the respective ecosystems (Miller et al., 2019). In such cases, understanding the pattern of their recurring nature can give insights into their groupings in both space and time. However, existing exploratory clustering methods to study ST event data consider a linear view of time, i.e., temporal proximity is defined based on events occurring within a specific time frame. This may not recognise the recurring nature of objects being studied (or study objects) that happen during a particular time period, for instance on a daily or weekly basis. In such cases, the nature of time needs to be viewed as cyclical. The current study aims to address this research gap in unsupervised ST event clustering and extend its application to the study of recurring Spatio-temporal groupings.

Another common limitation with the application of exploratory methods for clustering spatial or ST data acknowledging mobility is that the algorithms do not necessarily account for constraints offered by space. The use of Euclidian or planar distances over space overestimates the clusters(Lamb et al., 2016). Thus, the algorithms may not be able to form spatially proximate or contiguous clusters in the context of local topography. In such cases, there is a need to incorporate both - constraints to movement offered by physical networks and free movement over planar space. A graph-based representation of locations with ST events can help resolve this issue. The study aims to address this and apply it to the study of recurring Spatio-temporal groupings.

Also, grouping of ST events based on occurrence in space and time alone may not be adequate to explain the grouping. There is a need to consider the scale of time and space to identify clusters, i.e., at what distance must the event pairs be to be considered as near in space? Similarly, what should be the time frame to in which events need to be considered? (Lamb et al., 2020). Thus, in the context of identifying clusters based on ST groupings scale becomes important because, the results could have differences when considering clustering across different spatial and temporal scales within the same dataset.

The 'check-in' data from LBSN can aptly represent such a combination of space and time to study recurring ST groupings. This is because 'check-in' is allowed by such services only at recognised locations uniquely identified. As an LBSN data, Foursquare covers a large number of prominent locations in a city and has a significant userbase. Moreover, users' 'check-ins' at various locations are an expression of their movement in space-time. Hence such data offers the opportunity to study ST clustering of cyclic events.

The ability to explore and explain recurring Spatio-temporal groupings based on ST events can find varied applications. Following are some practical applications-

1. Researchers working in ecological studies, environmental agencies and organisations concerned with conservation can use such exploratory methods to identify clusters or hotspots of activity based on group behaviour of animals.
2. In epidemiology and public health, such exploratory methods can help public health authorities trace the potential populations that have been or can get infected due to disease outbreaks (Atluri et al., 2018; Shekhar et al., 2015) .
3. Urban planning agencies and transportation companies can use such exploratory methods to understand the movement behaviour in urban areas using data from remote sensing technologies (e.g., RFID tags, GPS data, Mobile networks, Beacons and others) for intelligent traffic management and designing urban space based on the knowledge of human activity (Assem et al., 2017; Yang et al., 2015).
4. In the domain of Business Intelligence – organising, categorising and targeting customers with similar Spatio-temporal behaviour can help develop better marketing and product placement strategies (Yang et al., 2015).

1.3. Research Objectives-

The objective of the research is to adopt a graph-based strategy to agglomerative hierarchical clustering framework for identifying spatio-temporal groupings considering recurring nature of ST events using appropriate spatial and temporal scales.

1.3.1. Sub-Objectives

Following are the sub-objectives associated with the research work-

a. To conceptualise a graph-based strategy for studying spatio-temporal grouping in agglomerative hierarchical clustering algorithm.
b. To identify recurrent nature of study objects in a given spatio-temporal event dataset by considering the cyclical nature of time.
c. To use appropriate spatial and temporal scales in clustering study objects for identifying meaningful spatio-temporal groupings.

d. To test the performance of the resulting clustering algorithm using a publicly available dataset.

1.3.2. Research Questions

1. Addressing functional requirements-

    a. How to incorporate a cyclical view of time in ST cluster? What would be an appropriate metric to capture it?
    b. How to decide the scale factors for determining spatial and temporal boundaries of observational pairs?
    c. What would be an appropriate method to visualise and interpret the results?

2. For Testing and Validation of Framework:

    a. What parameters are most important in determining the performance of the proposed algorithm?
    b. Are the results from the framework useful and interpretable?

1.4. Thesis Structure

This study contributes to developing an unsupervised exploratory clustering method to detect recurring Spatio-temporal groupings in ST event data by accounting for constraints offered by physical space. A graph-based strategy has been adopted to agglomerative hierarchical clustering framework for identifying Spatio-temporal groupings considering their recurring nature. The method is applied to "check-in" data for a location-based social network (LBSN) service to identify clusters or hotspots of users' activity in a city.

The thesis document is organized as follows-

1. Chapter 2 provides a review of relevant literature synthesizing scientific and technological studies conducted regarding the types of Spatio-temporal data, existing clustering algorithms and location-based social networks are presented.
2. Chapter 3 explains the methodology followed in this study. It describes 'check-in' data for the city of New York by Foursquare, a popular LBSN service used for this study. The process used to model the spatial and temporal constraints using graph and its integration into the agglomerative hierarchical clustering framework is discussed.
3. Chapter 4 provides the results along with its analysis. The performance of the method, along with the limitations, are also explored.
4. Chapter 5 provides the conclusions from the study. Also, the possibility for future studies is discussed.

2. LITERATURE

2.1. Overview

This chapter presents a review of relevant literature conducted for the study. To understand the different aspects applied in this work, following sections are reviewed:

- Spatio-temporal Data – Explores the different kinds of spatio-temporal data recognised in literature, along with their characteristics.
- Clustering Techniques – Surveys relevant academic work of existing clustering algorithms. An overview of these algorithms along with a summary is presented. This is followed by a review of the clustering methods.
- Location Based Social Networks – An analysis of the LBSN service along with a brief description of Foursquare is presented.

2.2. Spatio-temporal (ST) Data

(Kisilevich et al., 2009) in their review discuss about Spatio-temporal data in detail. The ST data has the unique distinction of capturing both spatial and temporal information. Spatial information refers to the location over a geographical surface. This can be represented as an address, using a system of global or local coordinates. In comparison, Temporal information either captures the point in time or a period of time. This can be represented as year, month, day, hour, minute, second or a combination, depending on the granularity. The combined information of both space and time helps understand persistence patterns in a phenomenon and highlight unusual patterns over time. Kisilevich et al., (2009) give a pictorial representation of ST data that can broadly exist, as shown in *Figure 2.1*.

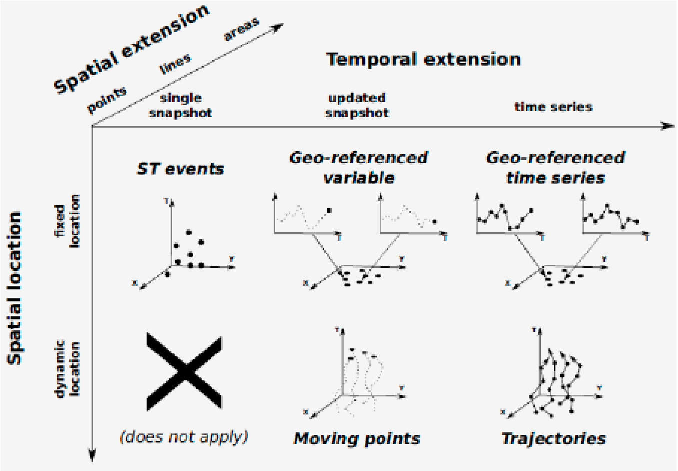


Figure 2.1 - *Types of spatio-temporal data (Source:*(Kisilevich et al., 2009)*)*

Spatio-Temporal (ST) Events: ST events are considered as elementary representations of Spatio-temporal information. An ST event contains the location and the corresponding timestamp of an event. This indicates where and when the event has been recorded. However, such information is considered static as no previous history related to either location or time is available. Thus, clustering ST events results in discovering close groups, both in space and time, and with other non-spatial properties (Kisilevich et al., 2009).

Geo-referenced variable: A geo-referenced variable enables one to observe the evolution of a phenomenon in time instances at a fixed location. Finding clusters, in this case, is similar to that of events, except that the Spatio-temporal objects compared at the locations have to be in the same instant of time, and the associated non-spatial features are not constant (Kisilevich et al., 2009).

Geo-referenced time series: A geo-referenced time series enables one to observe the entire history of a phenomenon at a fixed location. Finding clusters, in this case, involves comparing the way time-series evolve with spatial position. To achieve this, it is necessary to detect correlations between different time series in order to ensure that the effects of spatial autocorrelation are filtered out (Kisilevich et al., 2009).

Moving objects: A moving object is a case where the spatio-temporal data object's spatial location changes in time. As in the case of geo-referenced variables, the clustering challenge in this context consists of keeping the clusters updated. But in this case, incremental updating in results is required both in space and time (Kisilevich et al., 2009).

Trajectories: A trajectory represents a sequence of spatial locations visited by a spatio-temporal object along with the timestamps of the visits. Trajectories represent the behaviour in the movement of objects. Clustering, in this case, is used to identify groups of objects that behave similarly. It can be said that trajectories are a storehouse of the entire history of a moving object (Kisilevich et al., 2009).

Considering the different forms of ST data discussed above, the current study focuses on clustering ST events to discover spatio-temporal groupings. In ST events, as discussed above, the spatio-temporal information collected is considered static since no kind of evolution is possible. The following section reviews the state-of-the-art in clustering ST events.

2.3. Clustering Techniques for spatial and ST data

Clustering of data can be understood as a method to assess similarities in data and group them *(Figure 2.2)*. The objective of the assessment is to find new and interesting patterns from the data, which prima-facie may not be evident. These groupings can be defined in terms of proximity or attribute similarity, resulting in detection of different cluster types. Thus, (Birant & Kut, 2007) define clustering as grouping data based on similarity. A well-known famous application illustration of clustering to spatial data is that of Dr. John Snow, who in 1854 found clusters of cholera cases occurring around a public water pump which turned out to be the source for the spread of cholera (Shiode et al., 2015).

The proliferation of location-based technologies and reduction in the cost of storing data has allowed for vast amounts of spatial and ST data collection. This has made cluster analysis a topic of interest in spatial data mining, or also known as geographic knowledge discovery.

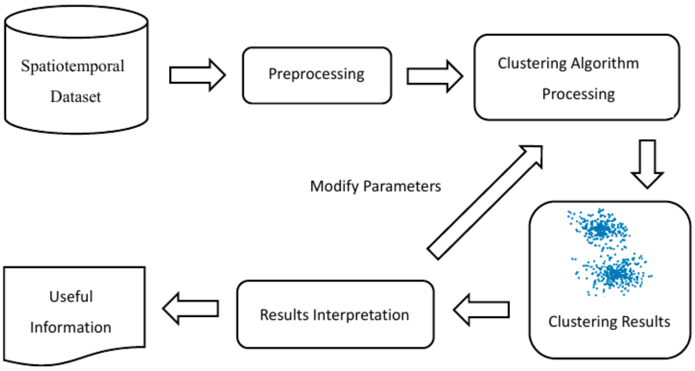


***Figure 2.2** – Schema for Clustering Procedure (Source*: (Shi & Pun-Cheng, 2019)*)*

The concept of 'cluster' is not precisely defined in the literature. This leads to a case where many algorithms are designed to achieve clustering; however, they may use the same or different methods (Estivill-Castro & Yang, 2000). Irrespective of the method used, clustering algorithms are expected to fulfil requirements such as (Marques, 2014):

1. Ability to deal with different types of attributes.
2. Ability to discover clusters with arbitrary shapes.
3. Ability to deal with noise and outliers.
4. Ability to incorporate domain knowledge in order to determine input parameters.
5. Ability to deal with multiple dimensions.
6. Interpretability of the results produced.
7. Usability.

The body of work associated with clustering of ST event data is based on the premise that some of the clustering algorithms for two-dimensional space can be generalized for spatiotemporal data (Han et al., 2010). Accordingly, the techniques for clustering spatial or spatio-temporal data are grouped as hypothesis testing-based and exploratory methods (Shi & Pun-Cheng, 2019).

2.3.1. Hypothesis testing-based methods for spatio-temporal clustering

Hypothesis based methods rely on statistical techniques and probability models for confirming an a priori hypothesis to identify the presence of clusters. Some known methods based on hypothesis testing for ST event clustering are discussed below(Shi & Pun-Cheng, 2019) .

Space-time interaction methods: The Knox and Mantel tests define what a closed group or a cluster means by providing a critical spatial and temporal distance manually. Those that satisfy the threshold are understood to have spatial and temporal adjacencies. Then the adjacencies are quantified between every two events separately.

Spatio-temporal k Nearest Neighbours Test: The k nearest neighbour test searches the k spatial and k temporal nearest neighbours for every event. The statistic looks at the space-time interactions that exist between event pairs when they are close in space and co-occur over time.

Scan Statistics: Scan statistics is a popular method for detecting clusters. The process uses a scanning window to look for similarities. A scan window can be defined with different extents to find clusters of two-dimensional spatial data with a statistical significance test since an appropriate extent is essential to detect meaningful clusters that are easy to interpret. Space and space-time scan statistics have similar calculation processes. Space-time scan statistics extend space scan statistics to detect clusters with the highest likelihood ratio by moving as a scan window covering the three dimensions to scan ST event data.

To summarize the hypothesis testing-based methods, those based on space-time interactions do not identify specific clusters. They are inefficient if the number of events is large. At the same time, those based on the Spatio-temporal k Nearest neighbour test are dependent on the value of 'k', which needs to be pre-defined. Moreover, a general drawback of hypothesis testing-based methods is that all of them are subject to the limitations of statistical methods.

2.3.2. Exploratory methods for clustering

The recent literature on spatial data mining provide the most widely accepted classification of spatial clustering methods into the following categories(Birant & Kut, 2007; Lamb et al., 2020; Shi & Pun-Cheng, 2019):

1. Partitioning methods

2. Hierarchical methods (agglomerative or divisive)
3. Density-based methods
4. Grid-based methods.

2.3.2.1. Partitioning Methods

Partitioning methods divide the data objects into exclusive groups (or clusters). The goal in such methods is to find a partition of k clusters wherein the chosen criterion for partitioning gets optimized and result in k clusters with the data objects uniquely associated to a cluster.

Thus, if a given data set, D has n objects, and if k is the required number of clusters to be formed, then the partitioning algorithm divides the set of objects into k partitions, such that each partition represents a cluster (Han et al., 2011).

The most common of all the partitioning methods is K-Means (Lloyd, 1982; MacQueen, 1967), which partitions data objects based on means. Based on this, a new algorithm was proposed by the name partitioning around medoids (PAM) or k-medoids (Leonard & Peter J., 1990). This was followed by clustering large applications (CLARA) to improve clustering efficiency("CLARA (Clustering LARge Applications)," 2009). Further clustering large applications based upon randomized search (CLARANS) was proposed to detect points and polygon objects (Ng & Han, 2002).

A drawback of the partitioning method of clustering is that it requires the number of desired output clusters pre-set by the user. This might be a problem, as the user may need to have good knowledge of the data to be clustered. Also, partitioning methods are not good with handling the discovery of clusters thar are arbitrary in shape. They are most suitable for concave spherical clusters. Lastly, they are susceptible to noise and outliers, and have difficulty in clustering data that contain them.

2.3.2.2. Hierarchical Methods

Hierarchical methods separate data objects into a hierarchy of levels based on a distance function (spatio-temporal). In other words, for a given data set D of n data objects, the measure of relative distances between the objects, based on a similarity criterion, is used to group the objects at different levels. The grouping involves a "sequence of irreversible steps" to construct a tree of clusters by either merging or splitting of data objects (Murtagh & Contreras, 2011) . The tree of clusters is referred to as a dendrogram. It helps the user decide the optimal number of clusters that can be formed from the given data. Unlike partition-based methods, the number of clusters need not be specified beforehand.

Hierarchical methods are categorized based on how the hierarchical decomposition is formed:

Agglomerative (bottom-up) – As the name suggests in this approach clustering starts with each object pair forming a separate group and then recursively merging into appropriate clusters based on distance. The process continues until a stopping criterion is met.

Divisive (top-down) – As the name suggests in this approach clustering starts with one cluster comprising of all the data objects. Then the single cluster recursively splits into smaller clusters. The process continues until a stopping criterion is met.

In the agglomerative approach, many methods have been proposed for merging clusters. These methods have been divided into two groups (Murtagh & Contreras, 2011):

- Group comprising of linkage methods- Single, Complete, Weighted, Unweighted

- Group comprising of methods which allow the cluster centres to be specified- Centroid, Median and minimum variance.

Depending upon the application either of the groups can be used. But the popular choice for merging clusters are the ones based on linkages.

Single linkage- This function defines the distance between the groups as the distance between the nearest cluster members of the respective groups.

Complete linkage- This function defines the distance between the groups by using the farthest possible distance between members of the respective groups.

Average group linkage- This function defines the distance between the groups by taking the average of the distances between members of respective groups. This method is also known as Unweighted Pair Group Method (UPGMA).

A drawback of the hierarchical methods of clustering in general is that the clusters once formed cannot be undone, i.e., it is not possible to revisit the constructed clusters for improvement improve (Murtagh & Contreras, 2011). Hence the resultant clusters may be of poor or low quality if the merge or split decisions, is not well chosen.

2.3.2.3. Density-based methods

As the name suggests, density-based methods clusters are obtained based on density. The density associated with a particular data object type can be understood as counting the number of similar points within a region with a specific radius (Ester et al., 1996). Thus, clusters are regions with a high density of data objects separated by regions where the density of data objects is low. These algorithms are popular for the purpose of database mining. Density-based methods are known for discovering clusters with arbitrary shapes. Like in the case of hierarchical clustering, the number of clusters need not be specified beforehand.

The commonly used density-based method is Density-based spatial clustering or DBSCAN, discussed above(Ester et al., 1996). OPTICS (Ordering Points to Identify the Clustering Structure) is another density-based clustering algorithm (Ankerst et al., 1999). Unlike DBSCAN, OPTICS computes the order in which data objects are clustered into the density-based structure. Lastly, DENCULE (DENsity-based CLUstEring) is a mixture of DBSCAN, Hierarchical Clustering and K-means (Hinneburg & Keim, 1998). The algorithm first estimates the local density of data objects in a way that is similar to kernel density estimation. Having estimated, clustered are defined as the local maximum for the estimated function.

By modifying DBSCAN to incorporate spatial and temporal information, Spatio-temporal DBSCAN or ST DBSCAN was proposed (Birant & Kut, 2007). ST-DBSCAN uses two filters to measure the similarities which is spatial values and non-spatial values. To incorporate the temporal aspect, the data objects are filtered by retaining only those data objects that are temporal neighbours. How temporal neighbours are considered is if the data objects occurred consecutively based on the timescale being considered, consecutive days, months or years (Birant & Kut, 2007).

A drawback of density-based clustering techniques is that they are sensitive to parametrization. This is applicable to ST-DBSCAN as well.

2.3.2.4. Grid-based methods

Grid-based methods divide the region where data objects are present into a grid. Clustering is then performed on the grid structure. This approach differs from the other exploratory methods for clustering spatial or ST data. Here clustering is not concerned with the actual data points but the space that surrounds the data point. the steps involved in grid-based clustering are described as follows (Grabusts & Borisov, 2002):

1. Creation of grid structure
2. Calculation of cell density for each cell
3. Sorting the cells by their cell densities
4. Identifying cluster centres

To study and analyse sequences of seismic events both spatially and temporally, ST-GRID algorithm was developed (Wang et al., 2006). The algorithm partitions spatial and temporal dimensions with varying precision over grid cells. This is followed by the extraction and merging of dense Spatio-temporal regions into clusters (Ansari et al., 2020). The algorithm utilises a neighbourhood search strategy and creates a graph for determining input parameters.

2.3.3. Review of Clustering methods

The following *Table 2.1* summarizes the exploratory clustering methods along with the popular algorithms.

Classification of Exploratory Clustering	Brief Account	Popular Algorithms
Partitioning	Data is partitioned into a user-specified number of groups. Each point belongs to one group. Does not work well for irregularly shaped clusters	k-means, k-medoids (PAM), CLARA, CLARANS.
Hierarchical	Data is arranged into a hierarchy of groups with the larger group containing the sub-groups. Two methods: agglomerative (builds groups from the observation up), or divisive (start with a large group and separate).	BIRCH, Chameleon, CURE
Density-based	Data is grouped based on a threshold for the number of objects in a neighbourhood. Useful for irregularly shaped clusters.	DBSCAN, OPTICS and DENCLUE
Grid-based	Data is grouped based on a grid structure, which is formed by dividing a region into a grid of cells.	STING, CLIQUE

Table 2.1 – Summary of Clustering methods and algorithms (Source: (Lamb et al., 2020))

A common limitation with the use of clustering algorithms listed in *Table 2.1* to spatial or ST data is that they may not consider the constraints offered by space. Use of Euclidian or planar distances over space overestimates the clusters (Lamb et al., 2016). Thus, the algorithms may not be able to form spatially proximate or contiguous clusters in the context of local topography.

Unlike regular clustering, ST clustering has distinct spatial boundaries and temporal boundaries. This needs to be integrated in a meaningful manner to form a cluster that accurately depicts the spatiotemporal phenomenon. This complicates the process. Incorporating both spatial and temporal information will require a strategy that identifies both spatial and temporal similarities and combines them (Kisilevich et al., 2009).The existing literature, adopt the following strategies to alleviate the problem of combining spatial and temporal information-

ST-DBSCAN filters ST data to retain successive temporal neighbours when calculating the distance for clustering ST objects. On the other hand, instead of calculating distance for clustering ST objects some of the approaches have adopted a neighbourhood search strategy and adapted it to existing algorithms. For example, adapting neighbourhood search strategy to ST-DBSCAN and ST-GRID (Wang et al., 2006), and to a spatio-temporal modification of OPTICS algorithm, i.e. ST-OPTICS (AgrawalK.P. et al., 2016). Carrying out Spatio-temporal clustering in a stepwise strategy or a particular order (spatial clustering followed by temporal clustering or vice-versa) has yielded results. For example, dividing the ST data discretely in accordance with the time period in which they occur and applying spatial clustering individually(Zhu & Guo, 2014). However, a drawback with such an approach is that clusters cannot be discovered across the discrete time periods. Though the solutions have yielded significant results, they were tested on trajectory data and on ST event data.

In some cases, space and time have been combined in a weighted manner to form a Spatio-temporal distance. For instance, transforming the temporal component into a spatial equivalent and then combining the two using Euclidean distance (AndrienkoGennady & AndrienkoNatalia, 2010)

Also, application depends on how the temporal distance and boundary between the two events need to be incorporated into clustering data. Existing ST clustering algorithms incorporate a linear view of time, i.e., temporal proximity is defined based on events occurring within a specific time frame. These algorithms do not recognise the recurring nature of ST data. The component dealing with temporal distance should consider it.

Irrespective of the strategies applied to combine spatial and temporal information in clustering there are certain requirements that have been determined for ST data in general (AgrawalK.P. et al., 2016). Firstly, the method used for clustering ST data should be able to determine clusters that are arbitrary or irregular in shape. Secondly, the method should have the ability to scale up and manage the high dimensionality of data. Thirdly, the method should be flexible enough to deal with spatial and temporal attributes such that the components can be added or removed as needed (AgrawalK.P. et al., 2016). Lastly, to be useful, the results should be interpretable.

2.4. Location Based Social Network

A location-based social network (LBSN) is a service that allows individuals in a social structure to share location embedded information. A social structure comprises of individuals who derive their interdependency- friendship, shared interests, and shared knowledge based on the content they share (such as video, photo and texts) from various locations in the physical world (Zheng, 2011). Thus, an LBSN service contributes to building and reflecting on the experiences of users over the Internet. There has been increased usage of LBSN with the proliferation of devices using location-based technologies (mobile phones, smart wearables, GPS etc.). With several people using such services, a large quantum of data is being generated, the nature of which is spatial, temporal or Spatio-temporal (ST) (Romero, 2011). With LBSN, users can add a spatial dimension to their posts on online social networks many ways. For instance, location-tagged media content (photos or videos) or share check-in details at a place or even share the trajectory of their trip.

(Zheng, 2011) categorises the applications providing LBSN services as: geo-tagged-media-based, point-location-driven and trajectory-centric services, the details of which have been discussed below.

Geo-tagged-media-based services: These services enable the users to incorporate a location label to their media content – which could be text, photos, and videos, while sharing it over the network. The ‘tagging’ is done either instantly or sometime later. This allows people to browse the content, either generated by them or shared over the network, with the location where it was created (Zheng, 2011). An example of such a service is Flickr which allows the user to share photos with the location.

Point-location-driven services: These services enable the users to share their current location through ‘check-in’ at venues. Sharing the real-time location allows the users to discover individuals (from their social network) around their location to augment social interactions and activities. For instance, they are inviting over people nearby for social activities (Zheng, 2011). Foursquare, Gowalla and Brightkite are examples of such services.

Trajectory-centric services: These services enable the users to capture both point locations and the route connecting the point locations or trajectory. Such services offer rich information about the users in the form of tags, tips, and photos over the trajectory and some basic information, such as distance, duration, and velocity (Zheng, 2011). Strava allows users to share their bicycle, jogging and other routes over the network along with the facility to add media content.

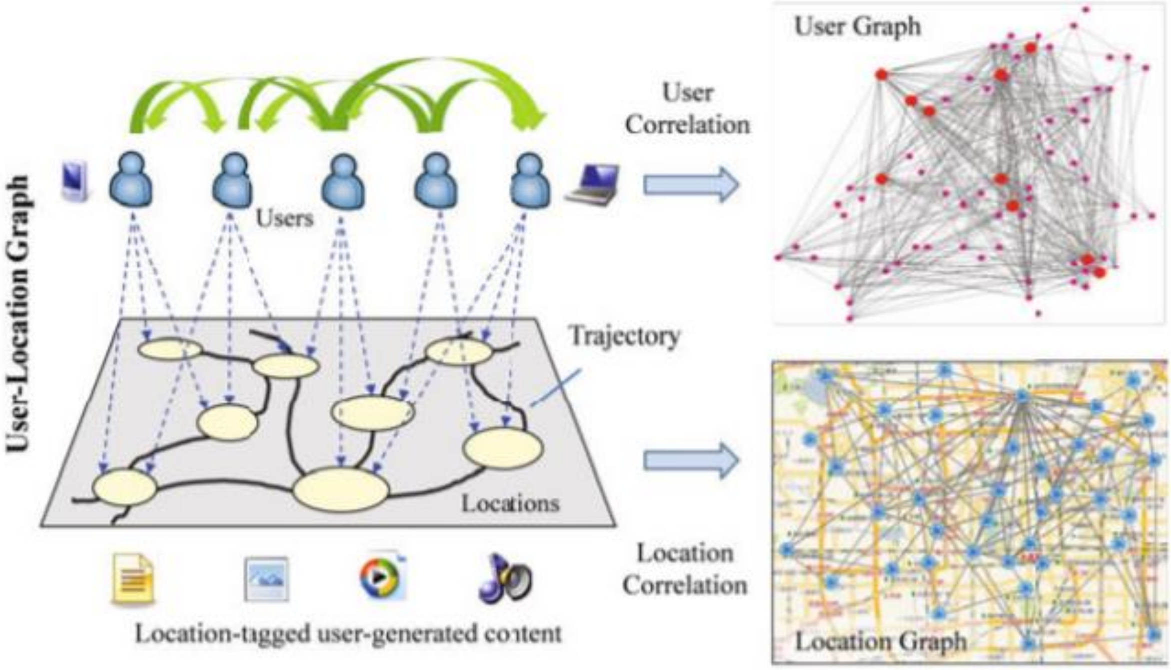


Figure 2.3 *– An image depicting possible research areas in LBSN (*Source: (Zheng, 2011)*)*

Figure 2.3 depicts the users' visits to various locations which accumulates to form their location histories along with location-tagged media content. Such information provides a lot of potential for understanding the users, locations, and relationship between them (Zheng, 2011).

LBSN data find use in several applications. From the point of view of locations, the data is an excellent source for discovering functional regions and discovering events in a given geography. LBSN data was to identify functional regions in urban areas (Assem et al., 2017). LBSN data was also used to discover sub-urban areas from foursquare data referred to as Livelihoods (Assem et al., 2017). From the point of view of users, LBSN data allows to estimate user similarity, discover communities, find local experts in a region, which allows for developing novel applications based on physical location (or activity). From the point of view of the relationship between user and location, enhanced recommendation systems and travel planning applications can be designed. LBSN data offers a plethora of opportunities in spatial science to study, for instance, social network analysis (modelling the user behaviour in accordance with location and time of visits and strength of connection between users based on spatio-temporal history), ST data mining, ubiquitous computing, and ST databases (Assem et al., 2017).

2.4.1. Foursquare

Founded in 2009, Foursquare is a location-based social network that engages users over its point-location-driven service platform by enabling them to share their location on a real-time basis through 'check-ins'. Users are virtually rewarded for engaging over the platform by sharing their locations or 'check-ins' into venues. It is estimated that Foursquare has over 105 million global points of interest contributed from over 500 million devices. The Foursquare application can be installed onto mobile devices. When at a place, they can check-in, letting their social network know where they are. Also, other users within the same network can see the checked-in location, which helps them to meet up. The check-ins can also be shared across other social networking platforms. The service allows the user to write reviews about the location. This is made available to other Foursquare users as well. Foursquare has a search engine- "Foursquare Explore" which provides personalized recommendations based on check-ins and reviews by other users for places in the user's vicinity. Also, from businesses' point of view, the platform acts as a medium for advertising to attract new customers and engage with existing customers. Other app developers actively use the Foursquare API to add a location to their apps. The application is available across different mobile operating system platforms - iOS, Android, BlackBerry, Windows Phone and other smartphones.

2.5. Visualization of time-space

Torsten Hägerstrand is attributed to the development of the concept of time-space geography in the mid-1960s. It was based on the research he had done on migration patterns of humans in Sweden. Time-space geography can be understood as a conceptual framework that incorporates the way individuals or groups allocate time and space. This helps in a trans-disciplinary understanding of spatial and temporal processes (Thrift, 1977) and gives an analytical perspective on movement and the associated activity pattern of humans in space-time (Annaler & Kwan, 2004). Also, studying the manner in which ST events occur in a time-space framework enhances our understanding of socio-environmental mechanisms. Thus, time-space geography finds application in transportation, planning, environmental science and ecology, and public health.

Though Hägerstrand is credited for developing the concept of time-space geography in the 1960s, the lack of means and technology did not allow it to be used to its fullest potential. The era before Geographic Information Systems (GIS) primarily relied on manual effort to compose paper maps and compute statistics to study the geospatial data. With time, analytical and map use techniques were developed to work with geospatial data, among them the concepts of time-geography. Today these techniques can be found in many GIS packages or developed to offer researchers and businesses access to powerful techniques to support their investigations.

As a part of time-space geography, Hägerstrand suggested a three-dimensional diagram to visualise how individuals or groups interact in space and time. The diagram also depicted the histories of the phenomenon as well as individuals or groups under consideration. The depiction of time as an addition to spatial dimension came to be known as a space-time cube. A typical space-time cube is the representation of a time-space phenomenon in three-dimensional space. The cube's base, with two of its dimensions, represents space, while the height of the cube represents the time dimension.

ST event analysis and visualisation involves looking for Spatio-temporal patterns in the data. This requires a holistic view of events in both space and time. Hence, there is a need to represent the complete space-time continuum along with the positions of the events in space as well as time, in the continuum. The idea of a space-time cube fits this need perfectly.

The usefulness of space-time cube for visualisation of Spatio-temporal patterns of events was favourably concluded(Gatalsky et al., 2004). Using the example of earthquakes at Maramara in Western Turkey, they demonstrate the effectiveness of space-time cube in detecting sequences of events that occurred in proximity with regard to space and time *(Figure 2.4)*.

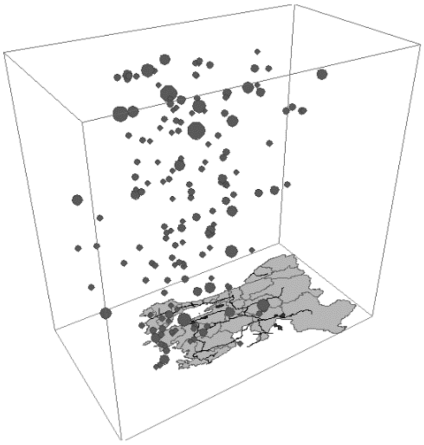


Figure 2.4 *– Space time cube for ST event data (Source:*(Gatalsky et al., 2004)*)*

Thus, Hagerstrand's concept of time-space helps in understanding and visualizing any phenomena characterised by spatial and temporal processes.

3. METHODOLOGY

3.1. Overview-

To develop an unsupervised exploratory method for detecting recurring Spatio-temporal groupings in ST event data, a graph-based strategy has been adopted to agglomerative hierarchical clustering framework. The method is applied to "check-in" data for a location based social network (LBSN) service to identify clusters or hotspots of activity of users in a city.

The approach considers a graph-based representation of ST events in the form of 'check-ins' which record a user's location and time. This approximates the ST events to a physical network and avoids the use of planar distances which may result in overestimation of clusters. An agglomerative hierarchical clustering framework is implemented to cluster 'check-in' events from the graph structure. Hierarchical clustering does not require pre-selection of clusters. This allows for automatic discovery of users based on the spatio-temporal similarity in check-ins. A flexible approach is adopted to identify spatio-temporal similarity. The use of a distance metric for clustering, *Equation 1* (detailed discussion in the following sections), where the spatial and temporal components are merged in a weighted linear combination with appropriate scales allows to conceive the spatial and temporal dimensions as required.

Equation 1

$$D(c_i, c_j) = \frac{1}{\sum_{i=1}^{n} w_i}\left(w_1 * \frac{D_s(x_i, x_j)}{S_D} + w_2 * \frac{D_t(t_i, t_j)}{S_T} + \sum other\ attributes\right)$$

To incorporate the notion of recurrence, a cyclic view of time has been adopted into the temporal distance. A brief workflow of the methodology has been depicted in *Figure 3.1*.

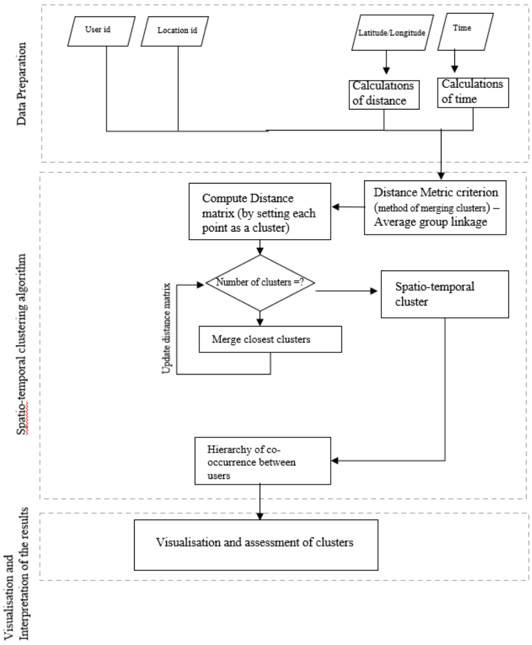

Data Preparation
User id
Location id
Latitude/Longitude
Time
Calculations of distance
Calculations of time
Spatio-temporal clustering algorithm
Compute Distance matrix (by setting each point as a cluster)
Distance Metric criterion (method of merging clusters) – Average group linkage
Number of clusters =?
Spatio-temporal cluster
Update distance matrix
Merge closest clusters
Hierarchy of co-occurrence between users
Visualisation and Interpretation of the results
Visualisation and assessment of clusters

Figure 3.1 *Workflow of the methodology*

The following sections discuss the methodology in detail.

3.2. Data-

Foursquare is a location-based social network that engages users over its point-location-driven service platform by enabling them to share their location on a real-time basis through 'check-ins'. It allows users to interact across the network by means of writing tips and comments, adding media content (photos or videos) about the place. The dataset used for this purpose is one that is publicly available (Yang et al., 2015). The dataset comprises of check-ins at various venues in the city of New York for a period of ten months (from 12 April 2012 to 16 February 2013). It contains a total of 227428 check-ins from 1083 users.

Data description:
The format of check-ins is in the form of tuple that consists of the following attributes:

1. User ID (anonymised)
2. Venue ID
3. Venue category ID
4. Venue category name
5. Latitude and Longitude
6. Time zone offset in minutes, UTC

A sample data looks as follows: {484, 4b5b981bf964a520900929e3, 4bf58dd8d48988d118951735, Food & Drink Shop, 40.69042712, -73.95468678, -240, Tue Apr 03 18:04:00 +0000 2012}.

Thus, for every given 'venue', the geographical coordinates are known. The information about the 'venue' is obtained from Foursquare users through crowdsourcing (Assem et al., 2017). Through this process, a 'venue' has been associated and classified accordingly with a semantic label that signifies a category to which it belongs. As noted previously, with regard to categories of places there are two broad categories that can be observed in Foursquare, a general abstract category (for instance, Arts and Entertainment on a broader note) and a specific category (for instance, Movie theatre under Arts and Entertainment) (Marques, 2014). In this study, the general categories are used. The specific categories can be clubbed into a total of nine general categories: Bar & Nightlife, Education, Food & Beverages, Work, Religious, Leisure & Outdoors, Shopping, Travel & Transport and Home. In *Table 3.1*, general category hierarchy and their respective specific categories considered can be seen. Also, in Figure the distribution of each abstract category in Foursquare dataset can be seen. There is a need to reduce the number of venue categories to a manageable number for better utility. Therefore, having a preliminary look at the categories, it was decided to reduce the number of venue categories from 251 to 9 *Figure 3.2* by grouping them based on the category hierarchy given by Foursquare.

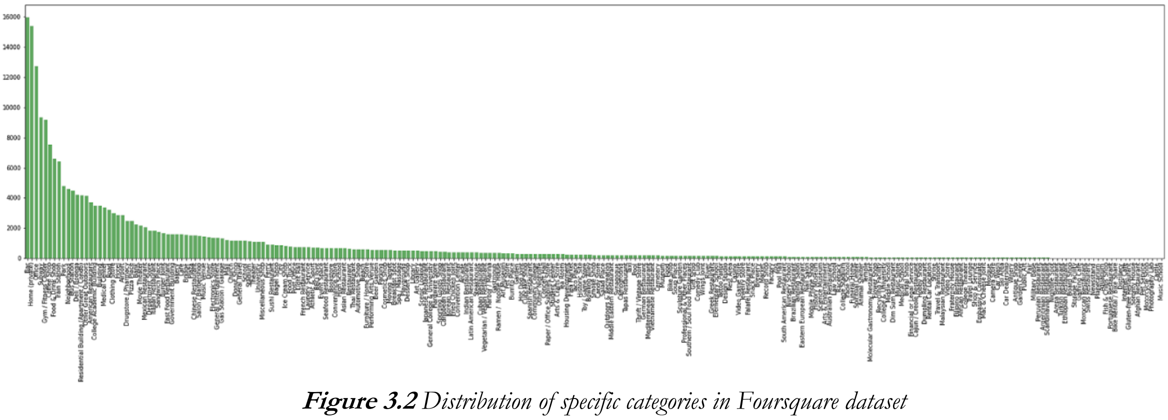


Figure 3.2 *Distribution of specific categories in Foursquare dataset*

Abstract Categories	Specific Categories
Bar & Nightlife	Gastropub, Bar, Comedy club, Casino, Beer Garden, Vinery, Brewery, Nightlife Spot, Distillery etc.
Food & Beverages	Eateries, Café, Bakery, Restaurant, Coffee Shop, etc.
Education	College, School, Fraternity House, Kindergarten, University, etc.
Religious	Church, Synagogue, Temple, Shrine, Mosque etc.
Leisure & Outdoors	Performing arts venue, Music venue, Art gallery, Scenic lookout, Campground, Athletics & Sports, Beach, Lake, Event space, Museum, Arcade, Park, etc.
Work	Government building, Office, Workplace, Post Office, Factory, Professional & other places, Embassy / Consulate, Financial & Legal Services etc.
Shopping	Arts & Crafts Store, Bookstore, Boutique, Clothing Store, Convenience Store, Electronic store, Mobile phone store, Furniture / Home Store, Mall, Supermarket, Plaza etc.
Travel & Transport	Ferry, Bus, Airport, Light rail, Gas Station, Bike share / Bike Rental, Taxi, Subway, Train Station, etc.
Home	Home (private), Residential Building, (Apartment/Condo), etc.

Table 3.1 *– Abstract and specific categories of Foursquare*

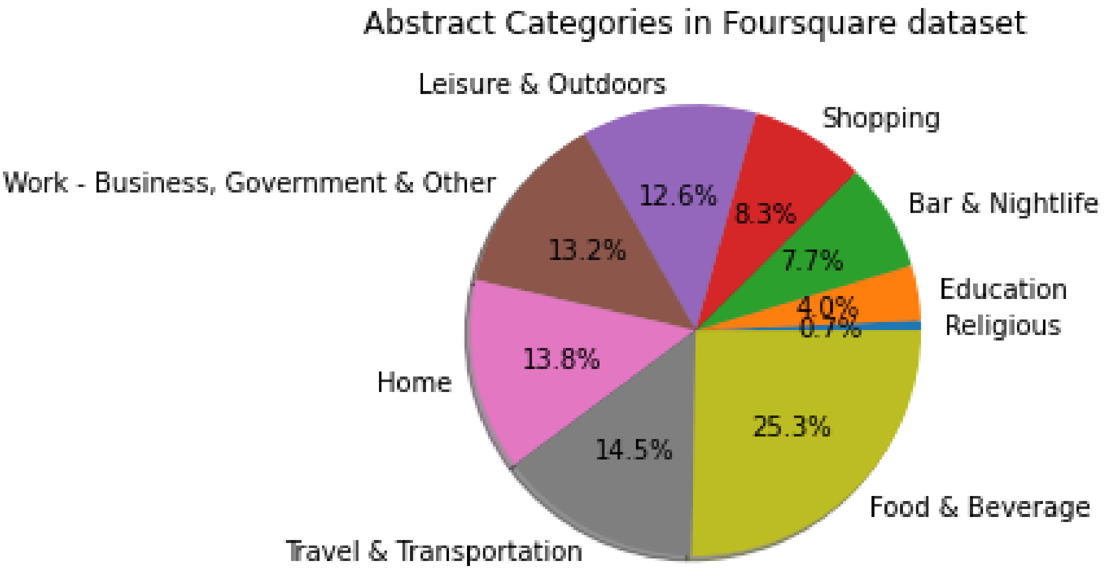


Figure 3.3 *Percentage distribution of each abstract categories in our Foursquare dataset.*

3.3. Concept of distance in space

ST event data, by definition, refers to point events represented by two dimensions in space and one dimension in time (Kisilevich et al., 2009). For clustering ST event data over space, the Euclidean distance or the planar distance between the events is usually considered, given by *Equation 2*.

Equation 2

$$D_E(x_i, x_j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

However, ST events associated with any kind of movement- be it trajectories, moving points or discrete events, are spatially influenced by the physical network in which the movements are observed. In this case, "check-in" data is a case of discrete ST events associated with users' movement. Thus, it is important to note that the physical network influences the movement through space in such cases. It was also observed that the use of planar spatial distance for clustering ST event points that are constrained by the physical network for movement overestimate the clustering (Lamb et al, 2015).

In reality, for clustering events that are in proximity over space and time, ST events associated with movement can neither be oversimplified by planar spatial distances nor will they strictly adhere to the physical network. For instance, two ST event points separated by a river may have to be connected by a route traversing through a bridge and not directly connecting the two points across the river as depicted in *Figure 3.4a*. On the other hand, navigation within buildings such as a shopping mall or office can be better represented by planar distances, as depicted in *Figure 3.4b*. Incorporating the constraints to movement offered by physical networks and free movement over planar space can help resolve this issue. A way to represent such a combination of spaces would be to determine a network of locations of ST events. Thus, a graph-based representation of locations where ST events are recorded can approximate movements to a physical network and thereby be used for clustering.

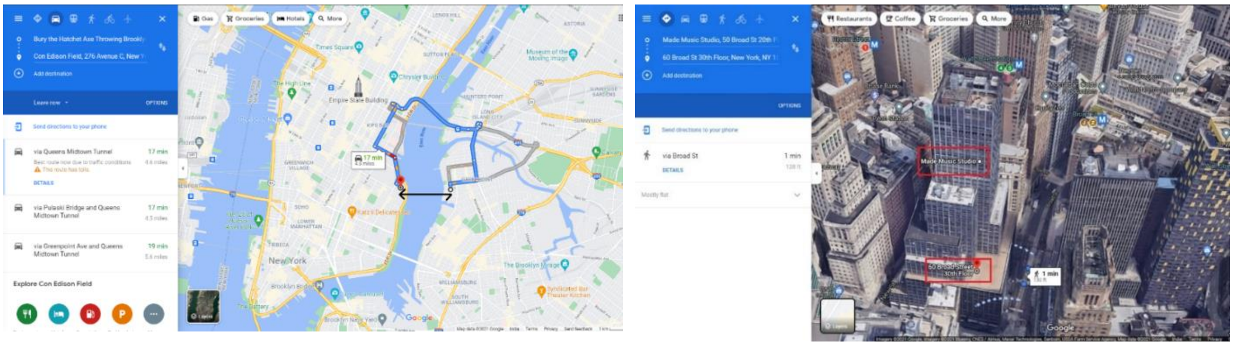

Bury the Hatchet Axe Throwing Brookl
Con Edison Field, 276 Avenue C, New Y
Add destination
Leave now
OPTIONS
Send directions to your phone
via Queens Midtown Tunnel
17 min
Best route now due to traffic conditions
This route has tolls.
DETAILS
via Pulaski Bridge and Queens Midtown Tunnel
17 min
via Greenpoint Ave and Queens Midtown Tunnel
19 min
Explore Con Edison Field
New York
Made Music Studio, 50 Broad St 20th
60 Broad St 30th Floor, New York, NY 1
Add destination
OPTIONS
Send directions to your phone
via Broad St
1 min
DETAILS
Mostly flat
1 min
Google

Figure 3.4 a. Issue with adhering to physical network to locate a point in space. b. Use of planar distances within the building (Source: New York City, Google Maps)

A neighbourhood graph could be way to represent locations. It is developed using Delaunay Triangulation (DT). The method tessellates a given space into a mesh of connected triangles. This helps mimic users' movement through places constrained by a physical network such as streets and where planar distance is applicable, such as shopping malls or buildings provided all the locations are known. A LBSN data in this context could be adopted to approximate the user's movement to a physical network. It also helps in adhering to the local topography of the place as depicted in *Figure 3.4 a.*, which may get ignored if only planar distances are considered. This could result in incorrect clustering of ST events over space. The following sub-sections discuss the formation of neighbourhood graphs.

3.3.1. Delaunay Triangulations

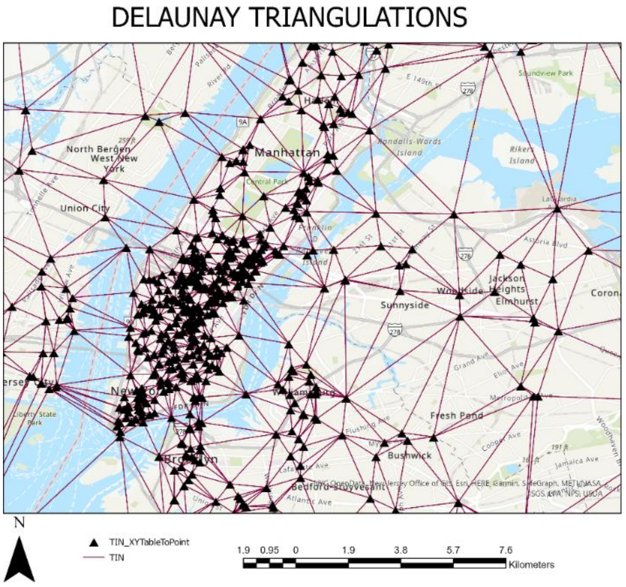


Figure 3.5 *Delaunay Triangulations for a combination of constraints to movement offered by physical networks and freedom of movement over planar space*

Creation of Delaunay triangulation marks the beginning of clustering. The vertices are defined by the 'check-in" events. Delaunay triangulation can be understood as a set of discrete over a Euclidean plane such that 'the circle circumscribing any three points in the triangulation contains no point of the same set inside it' (Lee & Schachter, 1980). Hence, Delaunay Triangulation helps present a data structure that looks for the closest neighbour to each 'check-in' event to establish a pairwise relation as seen in *Figure 3.5*. The pairwise relationships between the event locations established from the Delaunay Triangulation are used in the following section to create Graph data structure (Sibolla et al., 2021).

3.3.2. Graph-based Structure of 'Check-in' events

A graph is defined as "an ordered triple consisting of vertices, edges and an incidence function that associates the edges to the vertices "(Bondy & Murty, 1976). It is useful in building and finding relations between components that are discrete in nature for instance, people represented as points can be joined by lines based on whether they are friends or not. Hence it can be based on an attribute that describes the relation between the vertices. Since the idea of clustering is to establish and find relationships between ST data, concept of graph was modified to include the spatio-temporal distance as the relation between event points this study. The structure of a graph comprises of edges and nodes (or vertices) (Bondy & Murty, 1976). Nodes are connected to edges based on the attribute considered. The edge connects two nodes, thereby forming pairwise relationship. From the Delaunay triangulation, neighbourhood-based graph structure is crested using the vertices of the triangulations as nodes and the edges of the triangulations as graph edges (Sibolla et al., 2021).

The graph comprises of ST event nodes which are connected by edges. To compute the distance between the events the shortest path between the events was computed using Dijkstra's Algorithm (Xu et al., 2007). After calculating the distance, the distance was scaled $S_{D.}$

3.3.3. Nearest Neighbourhood Distance

To scale the distance component either the maximum distance between all the event points in the network can be considered or an adaptable distance which changes for each cluster point c_i could be used. The latter was chosen for this purpose as the clusters would be irregular in shape and would adapt depending on distance of c_i from other clusters.

For the selection of number of nearest neighbours *k,* local knowledge of the data would be useful. However, it was suggested to use the natural logarithm of the observations to get a value of *k* to start with (Birant & Kut, 2007).

3.4. Concept of distance in time

Distance in time depends on how one views time. Time can be viewed in terms of temporal proximity or cyclical occurrence. In the former case, interest in time is primarily to understand if events have occurred together or not. While in the latter case, interest in time is to understand the time of the day in which events have occurred. In this case events might not be occurring at regular intervals, but they occur during the same time of the day. Understanding recurring nature of events can give newer insights about temporal patterns in both space and time. Existing ST clustering algorithms incorporate a linear view of time, i.e., temporal proximity is defined based on events occurring within a specific time frame.

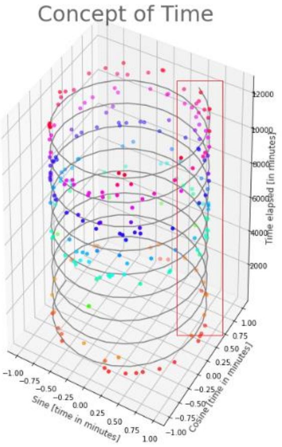


Figure 3.6 *Check-in activity of individuals in the morning hours at workplace for a period of 10 days.*

To understand the recurring nature of events, the time stamp of the events should be placed along the same timeline. For instance, *Figure 3.6* depicts two individuals 'checking-in' at office together in the morning will not only be temporally proximate but are checking in during the same time of the day which is morning.

To achieve this the difference in time between the check-in events as seen in *Equation 3,* is transformed to fit a cosine curve as shown in *Equation 4 and Equation 5.*

Equation 3

$$\Delta t = abs(t_{ic} - t_{jc})$$

Equation 4

$$t_{ij} = cos\left(\pi * \frac{\Delta t}{T}\right)$$

Equation 5

$$D_t(t_i, t_j) = \sqrt{(t_{ic}^2 + t_{jc}^2) - (2 * t_{ic} * t_{jc} * t_{ij}}$$

The time of event check-in is considered in minutes from mid-night T. The results t_{ij} vary between zero and one when scaled to radians. The Euclidian distance $D_{t,}$ between time stamps is computed using the law of cosine as seen in *Equation 5.*

3.5. Concept of scale

Having calculated the distance for ST event data, there is a need to scale the distance to obtain appropriate clusters. Firstly, scaling would help eliminate units. This allows for combining space and time into a common metric for computing distance between ST events which otherwise would not have been possible. Secondly, it helps in accounting for spatial and temporal boundaries thereby generating clusters with contiguous points. For instance, the scale factor in spatial context is to understand 'How far apart an event pair shall be considered as near to each other?'. The scale factor in temporal context is ensure that the temporal values are constrained to fit similar values within a time frame such as events separated by less than an hour and occurring together.

3.5.1. Scaling Spatial Distance

As discussed in *section 4.2.3,* for scaling the spatial distance, a varying distance for each ST event was used. The varying distance was based on the maximum distance of k nearest neighbours of ST event point under consideration. This allows for deciding the spatial extent to be considered for the ST events to fall within the same cluster.

3.5.2. Scaling Temporal Distance

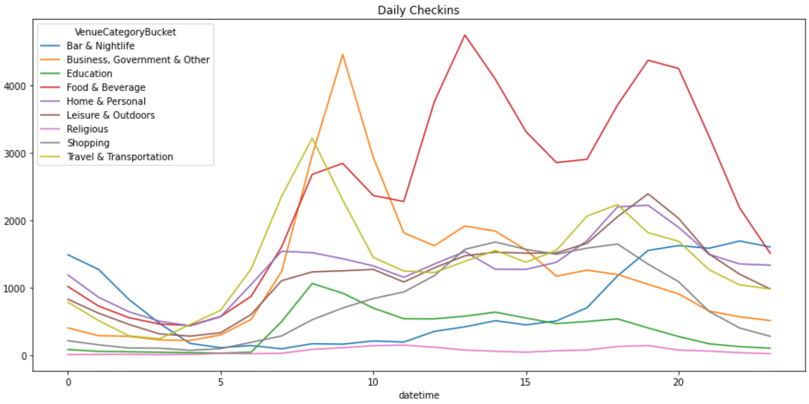


Figure 3.7 *Daily check-in counts for various abstract categories*

For scaling temporal distance, good knowledge of the data would help decide appropriate scale for the data.

For instance, to understand the dispersion of check-ins in time for various abstract categories and to identify the recurring patterns, a look at the distribution of daily check-in patterns *Figure 3.7* along with the results of Fourier transform may give a good idea to decide the scale value to be considered. Fourier transform helps in converting timestamps from a function of time to a function of frequency. This helps in identifying the underlying periodic patterns, if any.

Since the data comprises activities of users across spatial and temporal domain over a period, it is safe to assume that there could be recurring pattern associated with 'check-in' events. In order to understand the periodicity of the events, 'check-in' timestamps are aggregated into evenly spaced bins of suitable duration. The Fourier transform method is used to generate spectrum from the aggregated bins that can be treated as timeseries. Having done so, the periodicities around certain bin frequencies can be extracted. This helps in deciding the scale factor to be considered for time, i.e., the temporal distance over which two events can be assumed to have occurred in proximity.

Exploratory analysis of check-ins-

Using Fourier Transform, timestamps were transformed from a function of time into a function of frequency. This was used to identify the underlying periodic patterns by transforming time into the frequency domain.

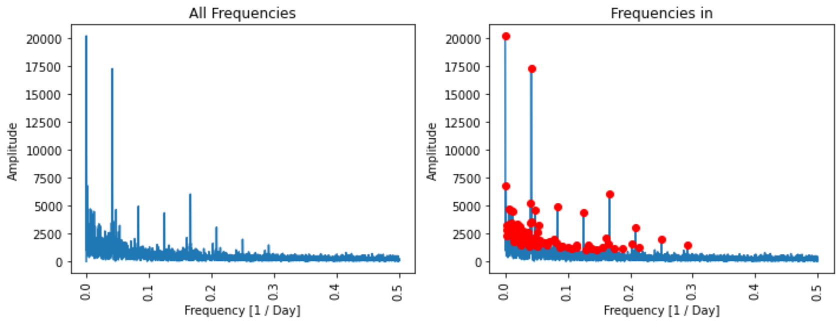


Figure 3.8 *Fourier Transform for each abstract category*

The raw data comprises count of number of check-ins - which can be considered for each hour or each day. For this assessment check-in counts are aggregated at the hour level and day level because check-in volume at the minute-level is too low and any periodicity below the hour-level is not expected. The frequencies with the highest amplitude are indicative of periodic patterns *Figure 3.8*. Frequencies with low amplitude are noise.

From the results of Fourier Transform for each abstract category, converting those frequencies with the highest amplitudes into hours and days, depicts that the periodic pattern for (Pearson Coefficient – 0.65 to 0.8) -

- Food and Beverages category has a daily frequency (the period is ~1 day) and an hourly frequency of 6 hours.
- Travel has a daily frequency and an hourly frequency of 12 and 8 hrs.
- Work has a daily frequency and hourly frequencies of 12 hours and 8 hours [attributed to different timings for various workplaces]
- Home has a daily frequency and hourly frequencies of 12. [2 times in a single day. This can be attributed to getting out of homes and returning.]
- Bar and Nightlife has a daily frequency and weekly frequencies of 7 days and 3.5 days. [The former suggests there are regular check-ins on a daily basis. The latter suggests that check in volumes is significant for a period of 3.5 days and 7 days.]
- Shopping periodic pattern for shopping has a daily frequency
- Education category has a daily frequency and an hourly frequency of 6 hours [and 12 hours. [check-in volume spikes 2 times in a single day - differential hourly spikes can be attributed to different types of institutions - high school/ university/ Medical college etc.
- no periodicity observed in other categories

Thus, the results of Fourier transform suggest that check-ins for *'Food and Beverage'* category can be considered to study the recurring nature of check-in events. If the scale factor is constrained to six hours of check-in activity or less at '*Food and Beverage*' locations as observed, then difference in temporal distance between cluster pairs falling within six hours would have the scale factor less than one while those more than six hours would be greater than one.

3.6. Hierarchical Clustering

An agglomerative hierarchical clustering framework was selected for the purpose of ST clustering due to two advantages it offers:

1. Selecting the number of clusters beforehand, is not needed; this allows for automatic discovery of ST events based on the spatio-temporal similarity.
2. It incorporates the idea of hierarchical structure, thereby allowing to visualise clusters at different levels of time, space or both as depicted in *Figure 3.9*

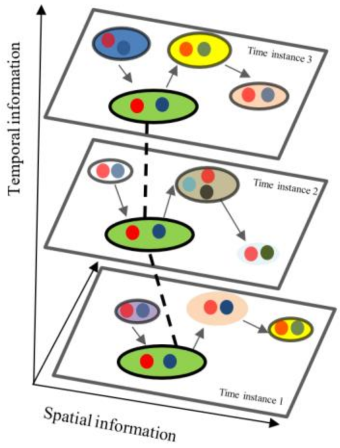


Figure 3.9 *Agglomerative hierarchical clustering framework*

The various components of *Equation 1* discussed in the preceding sections for location and time- Ds, D_t, S_D, S_T are combined in a linear combination with appropriate weightage as represented in *Equation 1*. This forms the function to compute the distance between the event pairs. The results of the function form the pairwise distance matrix of event clusters. Average linkage was used to develop clusters of events.

3.7. Execution

The graph-based strategy adopted to agglomerative hierarchical clustering framework approach was executed in Python 3.6 environment using the SciPy and NetworkX libraries. To obtain the graph structure, the Delaunay triangulations calculated for the 'check-in' event data that was created using the SciPy library. Subsequently Networkx library was used to obtain the graph-structure from the Delaunay triangulations and the network distance between 'check-in' event nodes in the graph was calculated using Dijkstra's algorithm to get the shortest path. This helped in computing the spatial distance and spatial scale for the 'check-in'. The fast Fourier transform (fft) algorithm offered by SciPy was used to compute the periodicity and thereby decide the scale factor. The distance function *Equation 1* created a distance matrix based on the inputs- Ds (*Equation 2*), D_t *(Equation 5),* S_D and S_T , with appropriate weights. The average linkage was used in developing the hierarchical clusters.

4. RESULTS AND DISCUSSION

4.1. Overview

To demonstrate its application in practice, the graph based agglomerative hierarchical clustering framework was used to study the "check-in" data by users of a LBSN service- Foursquare. The aim of the analysis was to identify clusters of activity by users that are recurring in both space and time within the city of New York.

4.2. Analysis

To practically demonstrate the application of graph-based strategy for agglomerative hierarchical clustering, a subset of check-in data of Foursquare for the city of New York was considered. The data comprised of top 50 uses, by check-in counts, whose check-in activity for a week's period starting from 5th May 2012 till 11th May 2012 was used. Further, the data was filtered to include 'check-ins' at venues categorised as 'Food and Beverages' (435 check-in events). This was based on the results of Fourier transform of discrete timestamps for 'check-ins' at various categories. Following explains the rationale:

1. To capture 'check-ins' that could show recurring nature - In order to capture 'check-in' events that have a recurring nature, it is necessary to understand the periodicity in 'check-in' patterns at various venues. Intuitively and otherwise, not all 'check-ins' of users can be treated in the same way as human movements are not random and have some purpose associated with them (Miller et al., 2019). Thus, the results of Fourier transform of discrete timestamps for 'check-ins' help in deciding which venues or category of venues can possibly show check-ins that are recurring daily.

   Note: In this study, recurring nature of 'check-in' events is being observed for different times of the day. This was decided based on the daily 'check-in' counts of the events for different times of the day. However, for studying recurrence on a weekly, fortnightly or monthly basis, the parameters for calculating temporal distance and scale factor have to be calibrated accordingly.

Venue Category	*Periodicity (day/hourly frequency)*
Food and Beverages	1 day/ less than 6 hours
Travel	1 day/12 hours/8 hours
Work	1 day/12 hours/8 hours
Home	1 day/12 hours
Bar and Nightlife	7 days / 3.5 days
Shopping	1 day
Education	12 hours/6hours
Leisure	No periodicity observed
Religious	No periodicity observed

   Table 4.1 *Results Fourier transform of discrete timestamps* for 'check-ins'

   The results help in deciding venue categories for which temporal clustering for different times of the day would be appropriate (Table – 4.1). In this case '*Food and Beverages*' was the only category that was deemed fit as it was the only category that could show results for recurrence of events during different times of the day.
2. To decide scale value for time that is comprehendible- The periodicity for the venue categories helps in deciding the scale factor to be considered for clustering. Since the scale factor for time is dependent on difference in temporal distance between cluster pairs, considering those cluster

pairs whose temporal difference is within a selected scale value would have the scale factor less than one. This would ensure that the temporal distance between the cluster pairs is kept low and the events would have higher chances of falling within the same cluster. Whereas a scale factor greater than one would contribute to increased temporal distance between the cluster pairs, diminishing the possibility of the event pairs falling within the same cluster.

Having accounted for temporal scale, for spatial scale, the natural log of the sample size (number of observations) was used to determine the spatial scale factor by computing the nearest neighbours. As pointed out, the rationale provided a '*k*' value to start with (Birant & Kut, 2007). A 'k' value of 6 and temporal scale of 60 minutes yielded some interesting results that are discussed below.

Figure 4.1 (a) shows the check-in counts of top 50 users at various Food and beverage outlets in the city for the said period. *Figure 4.1 (b)* and the following figures representing ST clusters presents an angle viewing from North-west towards South-east part of the city (to best show spatial and temporal information). The clustered results were analysed for different combinations of weights. Three cases were chosen to analyse the impact of weights on both spatial and temporal components of the clusters formed.

In the first case, the results were clustered with lower weightage being given to spatial distance component (0.1) and higher weightage being given to the temporal distance component (1.0). Thus, the figure emphasizes temporal characteristics of the 'check-in' events over spatial ones. The check-in activity at restaurants closely follows the trend of check-ins depicted by *Figure 4.1 (a)*. High check-in activity is observed around mid-day while it tapers down during early morning hours and late-night hours. The clustered results depict a similar trend when panned vertically. Also, horizontal slices depict the same colour indicating the recurring check-in behaviour for the duration, during said times of the day.

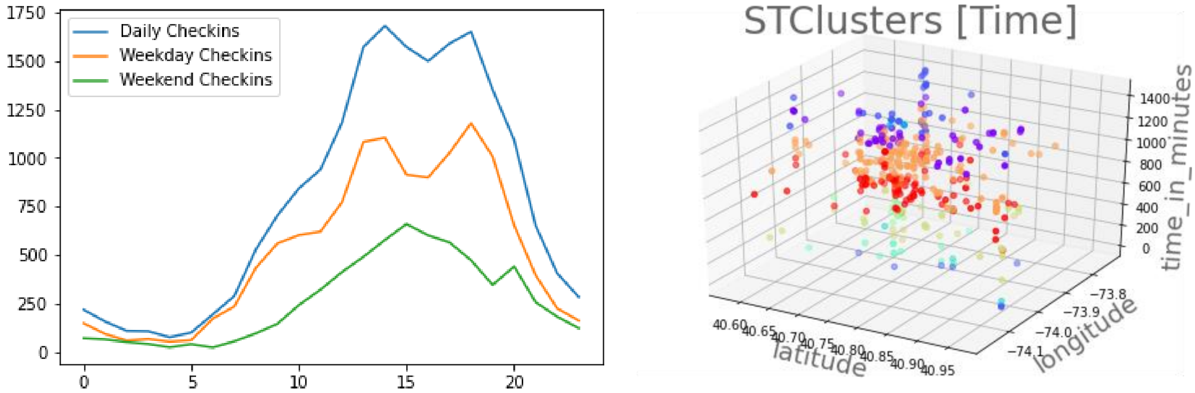


Figure 4.1 (a) Trend of check-in counts for the said period (b) Temporally clustered results for the said period

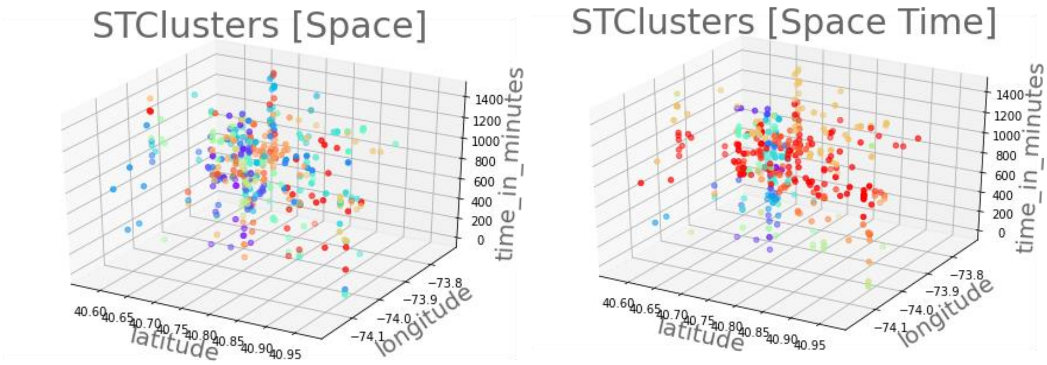


Figure 4.2 (a) Spatially clustered results for the said period (b) Spatio-Temporal clustered results for the said period

In the second case, the results were clustered with lower weightage being given to temporal distance component (0.1) and higher weightage being given to the spatial distance component (1.0). *Figure 4.2 (a)* depicts the results for the same value of *k* with lower weightage being given to time component (0.1) and higher weightage being given to the spatial component (1.0). Thus, the figure emphasizes spatial characteristics of the 'check-in' events over temporal ones. The clusters depict areas with highest number of visits under 'Food and Beverages' category, spread across different parts of the day for the time period. The northern and eastern part (area surrounding Yonkers, Bronx and Queen) show lower check-in volume whereas the southern part (Lower Manhattan area) shows high check-in volume. The clustered results depict a similar trend when panned horizontally. Also, vertical slices depict the same colour indicating the recurring check-in behaviour for the duration, one week, during said times of the day.

In the third case, the results were clustered with equal weightage to both spatial and temporal distance component (1.0). *Figure 4.2 (b)* depicts the result for spatio-temporal clusters with equal weightage being given to spatial and temporal components. The figure represents the venues in the city where recurring check-ins have been observed for the duration being considered. Another way of interpreting the results would be that the certain venues showcase routine check-ins by users and have high traffic during certain identified time of the day.

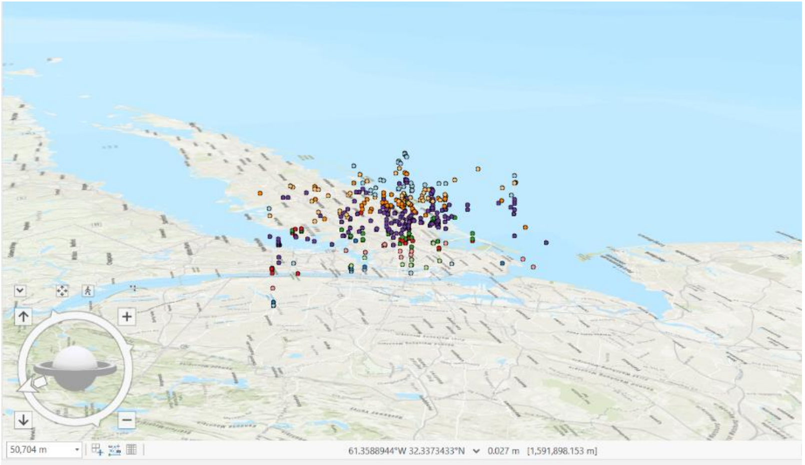


Figure 4.3 A 3D space-time visualization of check-in events

Figure 4.3 represents a space-time visualization of the check-in events.

4.3. Performance

The silhouette score computes the goodness of fit based on the inter-cluster distance and mean nearest-cluster distance. It indicates how close points in a cluster are to points in the neighbouring clusters (Rousseeuw, 1987). The value of silhouette score ranges from -1 to 1. A value of 1 indicates that a point in a cluster is far away from the neighbouring cluster. A value of 0 suggests that the point is close to the boundary of neighbouring cluster. While a value of -1 suggests that the point might have been wrongly assigned to a cluster. The silhouette score computation offered by scikit learn package was used to test the performance of the proposed method (Rousseeuw, 1987).

The silhouette score from *Table 4.2* suggests that two, three or four clusters would satisfactorily represent the data with k=3. While for k=6 it appears that only two clusters would fit well *Table 4.3*.

Method and Parameters	Cluster identified	Silhouette Score
Weightage – [1, 0.1] (importance to space over time)	2	0.727
	3	0.58
	4	0.443
Weightage – [0.1, 1] (importance to time over time)	2	0.661
	3	0.645
	4	0.392
Weightage – [1, 1] (importance to time and space)	2	0.674
	3	0.663
	4	0.406

Table 4.2 *Silhouette score Space scale - Nearest Neighbours, k = 3, Time scale 60 mins*

Space scale - Nearest Neighbours, k = 6, Time scale 60 mins		
Method and Parameters	Cluster identified	Silhouette Score
Weightage – [1, 0.1] (importance to space over time)	2	0.659
	3	0.403
	4	0.1
Weightage – [0.1, 1] (importance to time over time)	2	0.603
	3	0.34
	4	0.212
Weightage – [1, 1] (importance to time and space)	2	0.593
	3	0.309
	4	0.107

Table 4.3 *Silhouette score Space scale - Nearest Neighbours, k = 6, Time scale 60 mins*

The silhouette scores for clusters are satisfactory. However there seems to be scope for improvement with regard to quality of clusters.

4.4. Limitations

In the current work, a graph-based strategy has been adopted to an agglomerative hierarchical clustering framework. As noted previously there is a potential limitation with spatial and temporal components merged in a weighted linear combination (AndrienkoGennady & AndrienkoNatalia, 2010). It was noted that the interpretation of clusters might become problematic, as was seen in *Figure 5.2 (b)*. With the spatial and temporal components given equal weightage of one each, the clusters are challenging to interpret.

However, using a centrality measure in the graph can help overcome the issue by identifying meaningful clusters that can aid in the interpretation. Also, the use of statistical methods on observations associated with each cluster could help in interpretation.

The performance of the method can become a concern if used on large datasets. The reason being that hierarchical clustering produces a pairwise distance matrix to compute the linkages between data points. The distance matrix may need significant amount memory to handle the data.

Another limitation suggested in the literature is associated with the average linkage function used in agglomerative hierarchical clustering. The function computes merging decisions to form clusters. However, the merging is based on a static function (*see Equation 1*). This may lead to incorrect merging if the data is not cleaned correctly.

5. CONCLUSION AND FUTURE WORK

5.1. Conclusion

Recalling the sub-objectives stated for this study. The first was to conceptualise a graph-based strategy for studying spatio-temporal grouping in agglomerative hierarchical clustering framework. The proposed method was able to incorporate ST events that are network constrained and cannot be clustered using simple Euclidean or planar distances, into a graph for the purpose of hierarchical clustering. For this purpose, Foursquare data was used. The ST events were in the form of 'check-ins' by the users. The data is available publicly (Yang et al., 2015).

Second, there was the need to identify recurring nature of study objects in a given spatio-temporal event dataset by considering the cyclical nature of time. This was met by modifying the temporal distance component in the clustering algorithm. The difference in time between the 'check-in events' was transformed to fit a cosine curve. The method was useful in understanding recurring ST events occurring during times of the day.

Third, there was a need to incorporate appropriate spatial and temporal scales in clustering study objects for identifying meaningful spatio-temporal groupings. The scale factors incorporated, both spatial and temporal, allowed to do away with units for the purpose of combining spatial and temporal distance. It also helped understand clusters formed at different spatial and temporal scales. This helped in identifying meaningful spatio-temporal groupings along with retrieving additional knowledge with high relevance to location and time. Also, the allowance for weightage allowed for emphasizing on spatial component or temporal component in the data depending on the need.

Lastly, there was a need to test the performance of the clustering approach on broader parameters. Although the Silhouette score as a parameter was used to understand the goodness of fit of clusters based on the inter-cluster and mean-cluster distance. However, there is a need to test for other metrics such as adjusted random index- to measure the similarity of datapoints within the cluster and Homogeneity score for clusters. Also testing the performance of the stated method against similar established algorithms like ST-DBSCAN can be evaluated for recurring events by modifying the time component as stated in this study.

5.2. Future Work

Based on the noting of the current study, the following could be the areas of research for future work-

1. Need for comparative assessment between graph-based agglomerative hierarchical clustering approach and clustering considering Euclidean or planar distances-
   In theory, having incorporated ST events using a graph-based strategy into agglomerative hierarchical clustering framework, there were no metrics available to evaluate the nature of clusters formed that considers the constraints offered by local topography. An alternative could be to validate the results using ground truth data.

2. Computation of Spatio-temporal Distance
   The current approach utilises a linear combination of different components, in this case, spatial and temporal, in a weighted manner. However, the weights in the given function do not have any criteria and are randomly assigned to give relative importance to either spatial or temporal component. A criterion to define and understand weights and thereby incorporating into spatio-temporal clustering framework would allow for better interpretation of clusters.

3. Clustering Strategy
   The current approach combines the spatial and temporal component into a common distance metric. With the assigning of weights, this approach allows the user to compare the spatial and temporal groupings either in relation to one another (skewed weights) or independently (equal weights), but the combined distance by summing the spatial and the temporal distances lacks a reasonable explanation to comprehend the results quantitatively.

   An alternative to this could be to adopt a stepwise strategy of either spatial clustering followed by temporal clustering or vice-versa. This would yield meaningful clusters that could be rationally explained with the distances quantified.

4. Comparing performance with other ST clustering algorithms for event data-
   To test the efficacy of the approach a comparative assessment of performance and results with other similar algorithms such as ST-DBSCAN could be carried out to study the recurring nature of ST event data that are network constrained.

LIST OF REFERENCES

AgrawalK.P., GargSanjay, SharmaShashikant, & PatelPinkal. (2016). Development and validation of OPTICS based spatio-temporal clustering technique. *Information SciencesInformatics and Computer Science, Intelligent Systems, Applications: An International Journal*, *369*, 388–401. https://doi.org/10.1016/J.INS.2016.06.048

AndrienkoGennady, & AndrienkoNatalia. (2010). Interactive cluster analysis of diverse types of spatiotemporal data. *ACM SIGKDD Explorations Newsletter*, *11*(2), 19–28. https://doi.org/10.1145/1809400.1809405

Ankerst, M., Breunig, M. M., Kriegel, H.-P., & Sander, J. (1999). *OPTICS: Ordering Points To Identify the Clustering Structure*.

Annaler, G., & Kwan, M.-P. (2004). GIS METHODS IN TIME-GEOGRAPHIC RESEARCH GIS METHODS IN TIME-GEOGRAPHIC RESEARCH: GEOCOMPUTATION AND GEOVISUALIZATION OF HUMAN ACTIVITY PATTERNS. *Geogr. Ann*, 86–90.

Ansari, M. Y., Ahmad, A., Khan, S. S., Bhushan, G., & Mainuddin. (2020). Spatiotemporal clustering: a review. *Artificial Intelligence Review*, *53*(4), 2381–2423. https://doi.org/10.1007/s10462-019-09736-1

Assem, H., Xu, L., Buda, T. S., & O'Sullivan, D. (2017). *Spatio-Temporal Clustering Approach for Detecting Functional Regions in Cities*. 370–377. https://doi.org/10.1109/ICTAI.2016.0063

Atluri, G., Karpatne, A., & Kumar, V. (2018). Spatio-temporal data mining: A survey of problems and methods. *ACM Computing Surveys*, *51*(4). https://doi.org/10.1145/3161602

Birant, D., & Kut, A. (2007). ST-DBSCAN: An algorithm for clustering spatial–temporal data. *Data & Knowledge Engineering*, *60*(1), 208–221. https://doi.org/10.1016/J.DATAK.2006.01.013

Bondy, J. A., & Murty, U. S. R. (1976). *GRAPH THEORY WITH APPLICATIONS*.

CLARA (Clustering LARge Applications). (2009). In *Encyclopedia of Database Systems* (pp. 330–330). Springer, Boston, MA. https://doi.org/10.1007/978-0-387-39940-9_2177

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*. www.aaai.org

Estivill-Castro, V., & Yang, J. (2000). Fast and Robust General Purpose Clustering Algorithms. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *1886*, 208–218. https://doi.org/10.1007/3-540-44533-1_24

Gatalsky, P., Andrienko, N., & Andrienko, G. (2004). Interactive analysis of event data using space-time cube. *Proceedings of the International Conference on Information Visualization*, *8*, 145–152. https://doi.org/10.1109/IV.2004.1320137

Grabusts, P., & Borisov, A. (2002). *Using Grid-clustering Methods in Data Classification*.

Gudmundsson, J., Laube, P., & Wolle, T. (2008). Movement Patterns in Spatio-temporal Data. In S. Shekhar & H. Xiong (Eds.), *Encyclopedia of GIS* (pp. 726–732). Springer US. https://doi.org/10.1007/978-0-387-35973-1_823

Han, J., Kamber, M., & Tung, A. K. H. (2010). Spatial clustering methods in data mining. *Geographic Data Mining and Knowledge Discovery*, 188–217. https://doi.org/10.4324/9780203468029_CHAPTER_8

Hinneburg, A., & Keim, D. A. (1998). *An Efficient Approach to Clustering in Large Multimedia Databases with Noise*. www.aaai.org

Kisilevich, S., Mansmann, F., Nanni, M., & Rinzivillo, S. (2009). Spatio-temporal clustering. *Data Mining and Knowledge Discovery Handbook*, 855–874. https://doi.org/10.1007/978-0-387-09823-4_44

Lamb, D. S., Downs, J. A., & Lee, C. (2016). The network K-function in context: examining the effects of network structure on the network K-function. *Transactions in GIS*, *20*(3), 448–460. https://doi.org/10.1111/TGIS.12157

Lamb, D. S., Downs, J., & Reader, S. (2020). Space-Time Hierarchical Clustering for Identifying Clusters in Spatiotemporal Point Data. *ISPRS International Journal of Geo-Information 2020, Vol. 9, Page 85*, *9*(2), 85. https://doi.org/10.3390/IJGI9020085

Lee, D. T., & Schachter, B. J. (1980). Two algorithms for constructing a Delaunay triangulation. *International Journal of Computer & Information Sciences*, *9*(3), 219–242. https://doi.org/10.1007/BF00977785

Leonard, K., & Peter J., R. (1990). *Finding Groups in Data* (K. Leonard & P. J. Rousseeuw (eds.); First). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470316801.CH2

Lloyd, S. P. (1982). Least Squares Quantization in PCM. *IEEE Transactions on Information Theory*, *28*(2), 129–137. https://doi.org/10.1109/TIT.1982.1056489

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Https://Doi.Org/*, *5.1*, 281–298. 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/bsmsp/1200512992

Marques, F. (2014). A Constraint-Based Clustering Algorithm for Detection of Meaningful Places. *Undefined*.

Miller, H. J., Dodge, S., Miller, J., & Bohrer, G. (2019). Towards an integrated science of movement: converging research on animal movement ecology and human mobility science. *Https://Doi.Org/10.1080/13658816.2018.1564317*, *33*(5), 855–876. https://doi.org/10.1080/13658816.2018.1564317

Murtagh, F., & Contreras, P. (2011). *WIREs Data Mining Knowl Discov.* https://doi.org/10.1002/widm.53

Ng, R. T., & Han, J. (2002). CLARANS: A method for clustering objects for spatial data mining. *IEEE Transactions on Knowledge and Data Engineering*, *14*(5), 1003–1016. https://doi.org/10.1109/TKDE.2002.1033770

Oswaldo, A., & Romero, C. (2011). *Mining moving flock patterns in large spatio-temporal datasets using a frequent pattern mining approach*.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*(C), 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Shekhar, S., Jiang, Z., Ali, R. Y., Eftelioglu, E., Tang, X., Gunturi, V. M. V., & Zhou, X. (2015). Spatiotemporal Data Mining: A Computational Perspective. *ISPRS International Journal of Geo-Information 2015, Vol. 4, Pages 2306-2338*, *4*(4), 2306–2338. https://doi.org/10.3390/IJGI4042306

Shi, Z., & Pun-Cheng, L. S. C. (2019). Spatiotemporal Data Clustering: A Survey of Methods. *ISPRS International Journal of Geo-Information 2019, Vol. 8, Page 112*, *8*(3), 112. https://doi.org/10.3390/IJGI8030112

Shiode, N., Shiode, S., Rod-Thatcher, E., Rana, S., & Vinten-Johansen, P. (2015). The mortality rates and the space-time patterns of John Snow's cholera epidemic map. *International Journal of Health Geographics 2015 14:1*, *14*(1), 1–15. https://doi.org/10.1186/S12942-015-0011-Y

Sibolla, B. H., Van Zyl, T., & Coetzee, S. (2021). Determining Real-Time Patterns of Lightning Strikes from Sensor Observations. *Journal of Geovisualization and Spatial Analysis*, *5*(1), 4. https://doi.org/10.1007/s41651-020-00070-7

Thrift, N. J. (1977). *An introduction to time-geography*. Geo Abstracts, University of East Anglia.

Wang, M., Wang, A., & Li, A. (2006). Mining Spatial-temporal Clusters from Geo-databases. In X. Li, O. R. Zaïane, & Z. Li (Eds.), *Advanced Data Mining and Applications* (pp. 263–270). Springer Berlin Heidelberg.

Xu, M. H., Liu, Y. Q., Huang, Q. L., Zhang, Y. X., & Luan, G. F. (2007). An improved Dijkstra's shortest path algorithm for sparse network. *Applied Mathematics and Computation*, *185*(1), 247–254. https://doi.org/10.1016/J.AMC.2006.06.094

Yang, D., Zhang, D., Zheng, V. W., & Yu, Z. (2015). Modeling user activity preference by leveraging user spatial temporal characteristics in LBSNs. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *45*(1), 129–142. https://doi.org/10.1109/TSMC.2014.2327053

Zheng, Y. (2011). Computing with Spatial Trajectories. In *Computing with Spatial Trajectories* (pp. 243–276). Springer, New York, NY. https://doi.org/https://doi.org/10.1007/978-1-4614-1629-6_8

Zhu, X., & Guo, D. (2014). Mapping Large Spatial Flow Data with Hierarchical Clustering. *Transactions in GIS*, *18*(3), 421–435. https://doi.org/10.1111/TGIS.12100