Zhang, Dongzhi (2018) ***Periodic pattern mining from spatio-temporal trajectory data.*** **PhD Thesis, James Cook University.**

JAMES COOK UNIVERSITY

DOCTORAL THESIS

Periodic Pattern Mining from Spatio-temporal Trajectory Data

Author: Dongzhi ZHANG

A thesis submitted in fulfillment of the requirements
for the degree of Doctor of Philosophy

in the

Information Technology

December 14, 2018

Declaration of Authorship

I, Dongzhi ZHANG, declare that this thesis titled, "Periodic Pattern Mining from Spatio-temporal Trajectory Data" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: ____

Date: ____

Abstract

Dongzhi ZHANG

Periodic Pattern Mining from Spatio-temporal Trajectory Data

Rapid development in GPS tracking techniques produces a large number of spatio-temporal trajectory data. The analysis of these data provides us with a new opportunity to discover useful behavioural patterns. Spatio-temporal periodic pattern mining is employed to find temporal regularities for interesting places. Mining periodic patterns from spatio-temporal trajectories can reveal useful, important and valuable information about people's regular and recurrent movements and behaviours.

Previous studies have been proposed to extract people's regular and repeating movement behavior from spatio-temporal trajectories. These previous approaches can target three following issues, (1) long individual trajectory; (2) spatial fuzziness; and (3) temporal fuzziness. First, periodic pattern mining is different to other pattern mining, such as association rule ming and sequential pattern mining, periodic pattern mining requires a very long trajectory from an individual so that the regular period can be extracted from this long single trajectory, for example, one month or one year period. Second, spatial fuzziness shows although a moving object can regularly move along the similar route, it is impossible for it to appear at the exactly same location. For instance, Bob goes to work everyday, and although he can follow a similar path from home to his workplace, the same location cannot be repeated across different days. Third, temporal fuzziness shows that periodicity is complicated including partial time span and multiple interleaving periods. In reality, the period is partial, it is highly impossible to occur through the whole movement of the object. Alternatively, the moving object has only a few periods, such as a daily period for work, or yearly period for holidays.

However, it is insufficient to find effective periodic patterns considering these three issues only. This thesis aims to develop a new framework to extract more effective, understandable and meaningful periodic patterns by taking more features of spatio-temporal trajectories into account.

The first feature is trajectory sequence, GPS trajectory data is temporally ordered sequences of geolocation which can be represented as consecutive trajectory segments, where each entry in each trajectory segment is closely related to the previous sampled point (trajectory node) and the latter one, rather than being isolated. Existing approaches disregard the important

sequential nature of trajectory. Furthermore, they introduce both unwanted false positive reference spots and false negative reference spots.

The second feature is spatial and temporal aspects. GPS trajectory data can be presented as triple data (x, y, t), x and y represent longitude and latitude respectively whilst t shows corresponding time in this location. Obviously, spatial and temporal aspects are two key factors. Existing methods do not consider these two aspects together in periodic pattern mining.

Irregular time interval is the third feature of spatio-temporal trajectory. In reality, due to weather conditions, device malfunctions, or battery issues, the trajectory data are not always regularly sampled. Existing algorithms cannot deal with this issue but instead require a computationally expensive trajectory interpolation process, or it is assumed that trajectory is with regular time interval.

The fourth feature is hierarchy of space. Hierarchy is an inherent property of spatial data that can be expressed in different levels, such as a country includes many states, a shopping mall is comprised of many shops. Hierarchy of space can find more hidden and valuable periodic patterns. Existing studies do not consider this inherent property of trajectory.

Hidden background semantic information is the final feature. Aspatial semantic information is one of important features in spatio-temporal data, and it is embedded into the trajectory data. If the background semantic information is considered, more meaningful, understandable and useful periodic patterns can be extracted. However, existing methods do not consider the geographical information underlying trajectories.

In addition, at times we are interested in finding periodic patterns among trajectory paths rather than trajectory nodes for different applications. This means periodic patterns should be identified and detected against trajectory paths rather than trajectory nodes for some applications. Existing approaches for periodic pattern mining focus on trajectories nodes rather than paths.

To sum up, the aim of this thesis is to investigate solutions to these problems in periodic pattern mining in order to extract more meaningful, understandable periodic patterns. Each of three chapters addresses a different problem and then proposes adequate solutions to problems currently not addressed in existing studies. Finally, this thesis proposes a new framework to address all problems.

First, we investigated a path-based solution which can target trajectory sequence and spatio-temporal aspects. We proposed an algorithm called Traclus (spatio-temporal) which can take spatial and temporal aspects into account at the same time instead of only considering spatial aspect. The result indicated our method produced more effective periodic patterns based on trajectory paths than existing node-based methods using two real-world trajectories. In order to consider hierarchy of space, we investigated existing hierarchical clustering approaches to obtain hierarchical reference spots (trajectory paths) for periodic pattern mining. HDBSCAN is an incremental version of DBSCAN which is able to handle clusters with different densities to generate a hierarchical clustering result

using the single-linkage method, and then it automatically extracts clusters from a hierarchical tree. Thus, we modified traditional clustering method DBSCAN in Traclus (spatio-temporal) to HDBSCAN for extraction of hierarchical reference spots. The result is convincing, and reveals more periodic patterns than those of existing methods.

Second, we introduced a stop/move method to annotate each spatio-temporal entry with a semantic label, such as restaurant, university and hospital. This method can enrich a trajectory with background semantic information so that we can easily infer people's repeating behaviors. In addition, existing methods use interpolation to make trajectory regular and then apply Fourier transform and autocorrelation to automatically detect period for each reference spot. An increasing number of trajectory nodes leads to an exponential increase of running time. Thus, we employed Lomb-Scargle periodogram to detect period for each reference spot based on raw trajectory without requiring any interpolation method. The results showed our method outperformed existing approaches on effectiveness and efficiency based on two real datasets. For hierarchical aspect, we extended previous work to find hierarchical semantic periodic patterns by applying HDBSCAN. The results were promising.

Third, we apply our methodology to a case study, which reveals many interesting medical periodic patterns. These patterns can effectively explore human movement behaviors for positive medical outcomes.

To sum up, this research proposed a new framework to gradually target the problems that existing methods cannot handle. These include: how to consider trajectory sequence, how to consider spatial temporal aspects together, how to deal with trajectory with irregular time interval, how to consider hierarchy of space and how to extract semantic information behind trajectory. After addressing all these problems, the experimental results demonstrate that our method can find more understandable, meaningful and effective periodic patterns than existing approaches.

Acknowledgements

First and foremost, I would like to express my gratitude to my supervisors Professor Ickjai (Jai) Lee and Dr. Kyungmi (Joanne) Lee. They have been supportive since the days I began studying a Ph.D. I would like to thank them for encouraging my research and allowing me to complete my Ph.D. They helped me build on my academic knowledge, which will be invaluable for both my further research and my future career. During my Ph.D, each of my papers was carefully revised by my supervisors many times. When I faced difficulties with research, They helped me to solve these problems with a great deal of patience on each occasion. They are most responsible and patient supervisors. It is my honour to finish my Ph.D under their supervision.

I also take this opportunity to thank James Cook University for supporting me through its high quality research environment. I really enjoyed every moment in university.

I would also like to thank all my friends. They make my life much more colourful.

Special thanks to my family, my mother and my father who gave me the financial support so that I complete my Ph.D

I declare this thesis to these people with gratitude.

Statement of the Contribution of Others

Nature of Assistance	Names
Supervision support	Professor Ickjai Lee, and Dr Joanne Lee.
Research support	Professor Ickjai Lee, and Dr Joanne Lee.
Proofreading support	Professor Ickjai Lee, Dr Joanne Lee and Norall Mulder (paid external proofreader).
Financial support	James Cook University.

Contents

List of Figures

List of Tables

List of Abbreviations

Association Rule Mining	**ARM**
False Negative Rate	**FNR**
False Positive Rate	**FPR**
Fourier Transform and Autocorrelation	**FT&Auto**
Fuzzy Cognitive Maps	**FCMs**
Generalized Sequential Patterns	**GSP**
Global Positioning System	**GPS**
Global System for Mobile Communications networks	**GSM**
Hidden Markov Model	**HMM**
Periodic Pattern Mining	**PPM**
Radio Frequency IDentification	**RFID**
Reference Spot	**RS**
Sequential Pattern Mining	**SPM**
Spatial compactness	**SC**
Spatio-temporal compactness	**STC**
Temporal compactness	**TC**
Traclus(Spatio-temporal)	**Traclus (ST)**

Chapter 1

Introduction

In this chapter, we introduce the background of spatio-temporal trajectory data mining, motivation of our research, research content, structure and framework of this thesis.

1.1 Background

Collecting large-scale spatio-temporal trajectory data of moving objects has become quicker and easier due to the rapid development of location acquisition technologies such as GPS (Global Position System), RFID (Radio Frequency IDentification), GSM (Global System for Mobile Communications networks), geo-social network and Wi-Fi (Li et al., 2004; Hoyoung Jeung, 2014; Spinsanti et al., 2013). With the help of these techniques, various moving objects like vehicles, animals, human beings, and natural phenomena (such as cyclones, tornados and ocean currents) can be tracked. A spatio-temporal trajectory from a moving object is defined as a series of continuously sampled points (trajectory nodes) through which a moving object passes. The trajectory data can be represented as a temporal sequence of geographical locations. Figure 1.1 shows a spatio-temporal trajectory $\{(l_0, t_0), (l_1, t_1), (l_2, t_2), (l_3, t_3), (l_4, t_4), (l_5, t_5)\}$, where the blue solid circles represent sampled points, l_i is the geographical location for each sampled point, t_i is the time at sampled point, the blue line segments with arrows present the trace, and the arrow shows the direction of trace. The most common spatio-temporal trajectory data is GPS data, which can be generated from GPS-equipped devices such as smartphones, vehicles and GPS collars. GPS data are temporally ordered sequences of geolocation which are recorded by a built-in GPS device which is carried by a moving object. This thesis explores GPS trajectory data, which is generally classified according to its different applications:

1. Movement of people: people record their travels with smartphones. For instance, people share travel experiences with others (Zheng et al., 2011); athletes record their performance for analysis (Gudmundsson and Horton, 2017); Flickr, an online photo management and sharing application, attaches a location and timestamp to each photo (Lee, 2014);

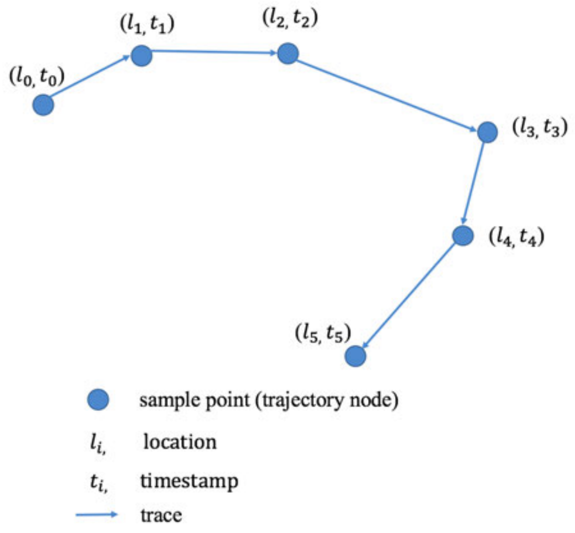


FIGURE 1.1: An example of spatio-temporal trajectory.

2. Movement of animals: GPS tracking collars are utilised to record animal′ trajectories in the study of animal′ migration and behaviors (Lee et al., 2007; Li et al., 2010a; Li et al., 2012; Bar-David, 2009);

3. Movement of vehicles: private and public transport are equipped with a GPS sensor that uploads real-time locations with a specific frequency for efficient traffic analysis (Gu et al., 2017; Yuan et al., 2013);

4. Movement of natural phenomena: Trajectories of some natural phenomena (such as cyclones, tornados and ocean currents) are collected to capture changes in climate. This enables people to better predict and manage natural disasters and to preserve the natural environment (Osuri, 2013; Mumby, 2011; Scott, 2017; Ashley, 2006).

Such a large number of long trajectories provides a new opportunity to discover meaningful, repetitive, and regular information and knowledge from the spatio-temporal context. However, this information cannot be directly extracted using traditional approaches. Periodic Pattern Mining (PPM) is a powerful technique that intelligently and automatically extracts regular and repetitive movement information from the vast amounts of spatio-temporal trajectory data (Cao et al., 2007; Li et al., 2012).

1.2 Motivation and Significance

A voluminous number of long movement trajectory data has been collected from various types of GPS-enabled tracking devices, such as built-in GPS mobile phones and PDAs, and sensors attached to animals. Spatio-temporal trajectory patterns can capture the movement patterns of objects which can plot the movements of objects in the study of their behaviours (Li et al., 2012; Zheng, 2015).

Periodicity is one of the common phenomena in moving objects. For instance, people go to work on weekdays and go shopping on the weekend, while animals migrate from one place to another. Both exhibit a certain periodicity. Thus, periodic patterns reveal repeated activities at regular time-intervals for a certain location. The mining of periodic patterns from spatio-temporal trajectories has attracted increased attention recently (Li et al., 2012). There is an increasing demand for its application for better understanding of the behaviours of moving objects (Cao et al., 2007; Li et al., 2012). For instance, knowledge of the periodic patterns exhibited by animals is crucial to their conservation. Likewise, the movements of motorists can assist congestion monitoring and traffic control. Furthermore, periodic patterns can provide valuable information to assist with decision-making (Li et al., 2010a; Han et al., 2010; Cao et al., 2007; L. Zhu et al., 2012; Jindal et al., 2013). Figure 1.2 shows an example of three interesting periodic patterns. John demonstrates three periodic patterns; periodic pattern 1→2 shows that he usually goes to university ② from home ① every day. Periodic pattern 2→3 displays that he goes to Smithfield shops ③ after school every week, and periodic pattern 1→4 demonstrates that he goes to Trinity Beach ④ from home ① on the weekend. This particular example shows how these periodic patterns can be used to analyse regular behaviours of moving objects. In addition, periodic patterns can also be used for compression of movement data (Cao et al., 2007; Mamoulis et al., 2004) and future movement prediction (Jeung et al., 2008; Li et al., 2012).

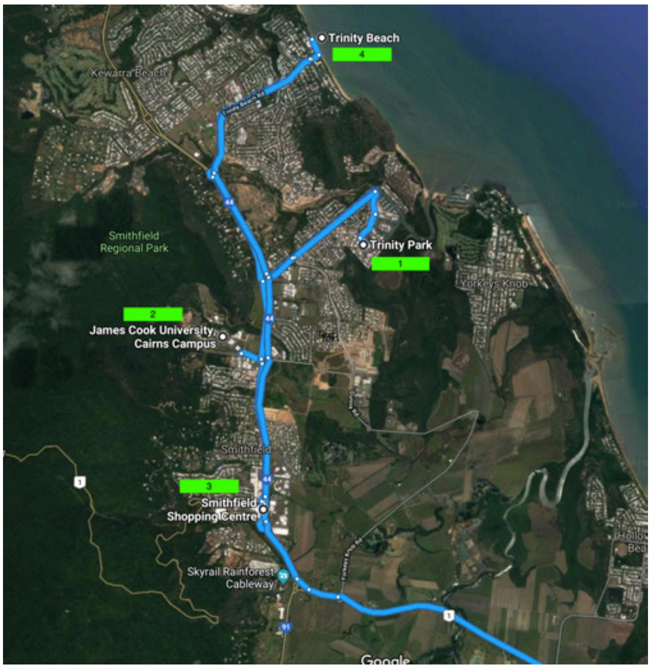


FIGURE 1.2: An example of periodic patterns.

In traditional periodic pattern mining, many studies have been conducted in the analysis of different types of datasets to obtain periodic patterns. These studies cover the following contexts:

- periodic pattern mining in event/symbol sequence data (Cao et al., 2004; Dong, 2009);
- periodic pattern mining in time series data (Jiawei Han, 1999; Aref et al., 2004; Yang et al., 2002);
- periodic pattern mining in social network data (Lahiri and Berger-Wolf, 2010; Lahiri and Berger-Wolf, 2008).

Approaches to these datasets cannot be directly applied to locating periodic patterns for spatio-temporal trajectories due to the unique characteristics of spatio-temporal trajectories: the uncertainty of spatio-temporal trajectories including locational (spatial) fuzziness, temporal fuzziness, the hierarchical nature of spatio-temporal phenomena, and irregularities of time intervals.

Also, existing spatio-temporal periodic pattern mining approaches suffer from some drawbacks (Cao et al., 2007; Mamoulis et al., 2004; Li et al., 2010a; Jindal et al., 2013; Li et al., 2012; Li et al., 2011):

- cannot consider sequence of trajectory;
  Figure 1.3 shows a spatio-temporal trajectory. This is a temporal sequence where each trajectory node is related to the previous one or the next one. For example, trajectory node (x_1, y_1, t_1) is related to trajectory node (x_2, y_2, t_2). Existing periodic pattern mining approaches failed to consider the sequence of trajectory.

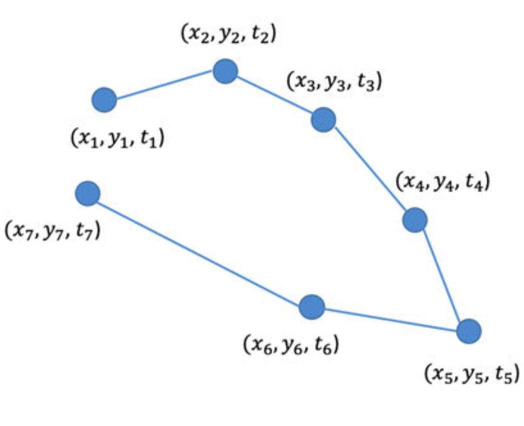
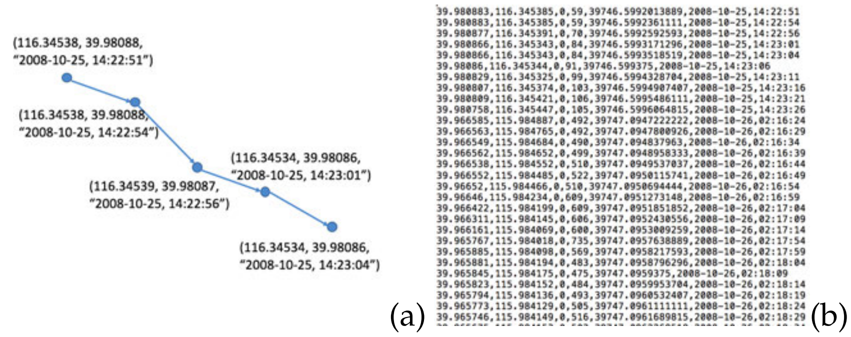


FIGURE 1.3: An example of sequence of trajectory.

- cannot deal with spatio-temporal trajectories with irregular time intervals;
  Figure 1.4(a) illustrates a spatio-temporal trajectory which is from Figure 1.4(b). Obviously, time intervals between two trajectory nodes are different. Existing periodic pattern mining approaches failed to

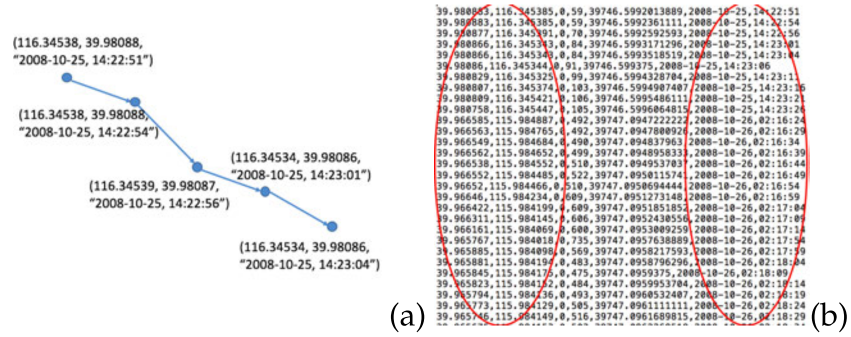


```
39.980883,116.345385,0,59,39746.5992013889,2008-10-25,14:22:51
39.980883,116.345385,0,59,39746.5992361111,2008-10-25,14:22:54
39.980877,116.345391,0,70,39746.5992592593,2008-10-25,14:22:56
39.980866,116.345343,0,84,39746.5993171296,2008-10-25,14:23:01
39.980866,116.345343,0,84,39746.5993518519,2008-10-25,14:23:04
39.98086,116.345344,0,91,39746.599375,2008-10-25,14:23:06
39.980829,116.345325,0,99,39746.5994328704,2008-10-25,14:23:11
39.980807,116.345374,0,103,39746.5994907407,2008-10-25,14:23:16
39.980809,116.345421,0,106,39746.5995486111,2008-10-25,14:23:21
39.980758,116.345447,0,105,39746.5996064815,2008-10-25,14:23:26
39.966585,115.984887,0,492,39747.0947222222,2008-10-26,02:16:24
39.966563,115.984765,0,492,39747.0947800926,2008-10-26,02:16:29
39.966549,115.984684,0,490,39747.094837963,2008-10-26,02:16:34
39.966562,115.984652,0,499,39747.0948958333,2008-10-26,02:16:39
39.966538,115.984552,0,510,39747.0949537037,2008-10-26,02:16:44
39.966552,115.984485,0,522,39747.0950115741,2008-10-26,02:16:49
39.96652,115.984466,0,510,39747.0950694444,2008-10-26,02:16:54
39.96646,115.984234,0,609,39747.0951273148,2008-10-26,02:16:59
39.966422,115.984199,0,609,39747.0951851852,2008-10-26,02:17:04
39.966311,115.984145,0,606,39747.0952430556,2008-10-26,02:17:09
39.966161,115.984069,0,600,39747.0953009259,2008-10-26,02:17:14
39.965767,115.984018,0,735,39747.0957638889,2008-10-26,02:17:54
39.965885,115.984098,0,569,39747.0958217593,2008-10-26,02:17:59
39.965881,115.984194,0,483,39747.0958796296,2008-10-26,02:18:04
39.965845,115.984175,0,475,39747.0959375,2008-10-26,02:18:09
39.965823,115.984152,0,484,39747.0959953704,2008-10-26,02:18:14
39.965794,115.984136,0,493,39747.0960532407,2008-10-26,02:18:19
39.965773,115.984129,0,505,39747.0961111111,2008-10-26,02:18:24
39.965746,115.984149,0,516,39747.0961689815,2008-10-26,02:18:29
```

FIGURE 1.4: An example of irregular trajectory.

extract periodic patterns from irregular spatio-temporal trajectory data.

- cannot consider spatial and temporal aspects at the same time;
  In Figure 1.5(a), spatio-temporal trajectory data are composed of triple data which include geographical location and timestamp (In Figure 1.5(b), the columns in red ellipse represent the geographical locations and timestamps). Spatial and temporal aspects are inherent properties for spatio-temporal trajectory data. Existing approaches to periodic pattern mining failed to consider spatio-temporal aspects at the same time in the process of periodic pattern mining.



```
39.980883,116.345385,0,59,39746.5992013889,2008-10-25,14:22:51
39.980883,116.345385,0,59,39746.5992361111,2008-10-25,14:22:54
39.980877,116.345391,0,70,39746.5992592593,2008-10-25,14:22:56
39.980866,116.345343,0,84,39746.5993171296,2008-10-25,14:23:01
39.980866,116.345343,0,84,39746.5993518519,2008-10-25,14:23:04
39.98086,116.345344,0,91,39746.599375,2008-10-25,14:23:06
39.980829,116.345325,0,99,39746.5994328704,2008-10-25,14:23:11
39.980807,116.345374,0,103,39746.5994907407,2008-10-25,14:23:16
39.980809,116.345421,0,106,39746.5995486111,2008-10-25,14:23:21
39.980758,116.345447,0,105,39746.5996064815,2008-10-25,14:23:26
39.966585,115.984887,0,492,39747.0947222222,2008-10-26,02:16:24
39.966563,115.984765,0,492,39747.0947800926,2008-10-26,02:16:29
39.966549,115.984684,0,490,39747.094837963,2008-10-26,02:16:34
39.966562,115.984652,0,499,39747.0948958333,2008-10-26,02:16:39
39.966538,115.984552,0,510,39747.0949537037,2008-10-26,02:16:44
39.966552,115.984485,0,522,39747.0950115741,2008-10-26,02:16:49
39.96652,115.984466,0,510,39747.0950694444,2008-10-26,02:16:54
39.96646,115.984234,0,609,39747.0951273148,2008-10-26,02:16:59
39.966422,115.984199,0,609,39747.0951851852,2008-10-26,02:17:04
39.966311,115.984145,0,606,39747.0952430556,2008-10-26,02:17:09
39.966161,115.984069,0,600,39747.0953009259,2008-10-26,02:17:14
39.965767,115.984018,0,735,39747.0957638889,2008-10-26,02:17:54
39.965885,115.984098,0,569,39747.0958217593,2008-10-26,02:17:59
39.965881,115.984194,0,483,39747.0958796296,2008-10-26,02:18:04
39.965845,115.984175,0,475,39747.0959375,2008-10-26,02:18:09
39.965823,115.984152,0,484,39747.0959953704,2008-10-26,02:18:14
39.965794,115.984136,0,493,39747.0960532407,2008-10-26,02:18:19
39.965773,115.984129,0,505,39747.0961111111,2008-10-26,02:18:24
39.965746,115.984149,0,516,39747.0961689815,2008-10-26,02:18:29
```

FIGURE 1.5: An example of spatial and temporal aspects.

- cannot take hierarchy of space into account;
  Spatial hierarchy is inherent in spatial data. For instance, in Figure 1.6, a moving object goes to James Cook University from home, and then goes to Smithfield Shopping Centre, and finally he/she returns home. Existing periodic pattern mining approaches only consider the single-level hierarchy. These approaches found periodic patterns between home, James Cook University and Smithfield shopping centre, but failed to find hierarchical periodic patterns between Smithfield area and Trinity Beach area.

- cannot consider background semantic information of spatio-temporal trajectories;

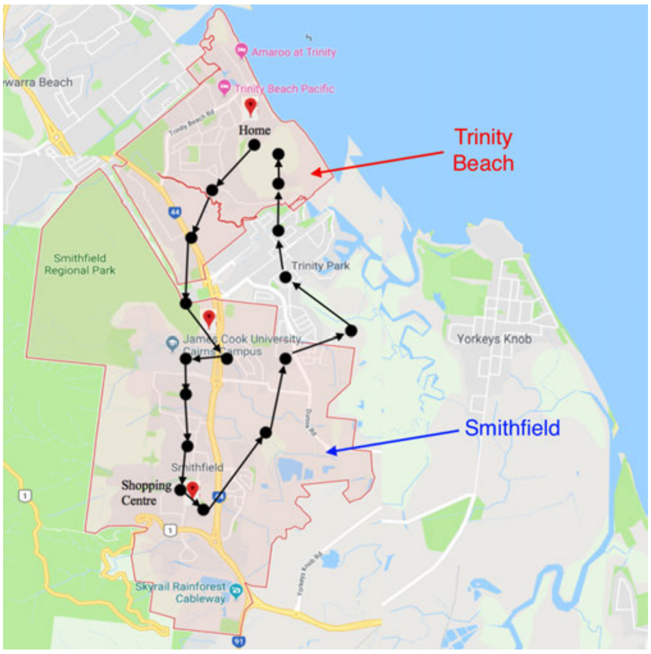


FIGURE 1.6: An example of hierarchy of space.

Background semantic information is very useful to infer people's purpose in moving to a particular place. Especially for periodic pattern mining, it is vital to know the aim of people's repeating and regular movement behaviors. For instance, In Figure 1.7(a), a moving object is moving around the space, but we do not know where he is going, when he is going and what is his purpose. In Figure 1.7(b), if we add a small amount of semantic background information, we can find he is moving around in the Trinity Beach area. If we add further semantic background information in Figure 1.7(c), we can see he goes to a gym, his workplace, his church, a shopping centre and a playground. Finally, we can find some periodic patterns, such as in Figure 1.7 (d), where he goes to his workplace at 9 am and then returns at 5 pm. Existing approaches cannot find these types of semantic periodic patterns.

- Existing periodic pattern approaches focus on trajectory nodes rather than paths;
  Sometimes, we are interested in finding some periodic patterns for trajectory paths based on particular needs, such as road planning. However, existing approaches focus on obtaining periodic patterns for trajectory nodes while ignoring trajectory paths. For instance, In Figure 1.8, existing methods can find the reference spots in blue circle areas but fail to find trajectory paths in red rectangle areas.

These limitations and drawbacks provide incentive to developing new approaches to periodic pattern mining from spatio-temporal trajectories. These six drawbacks of existing models are discussed in detail in Chapter 2 and Chapter 3

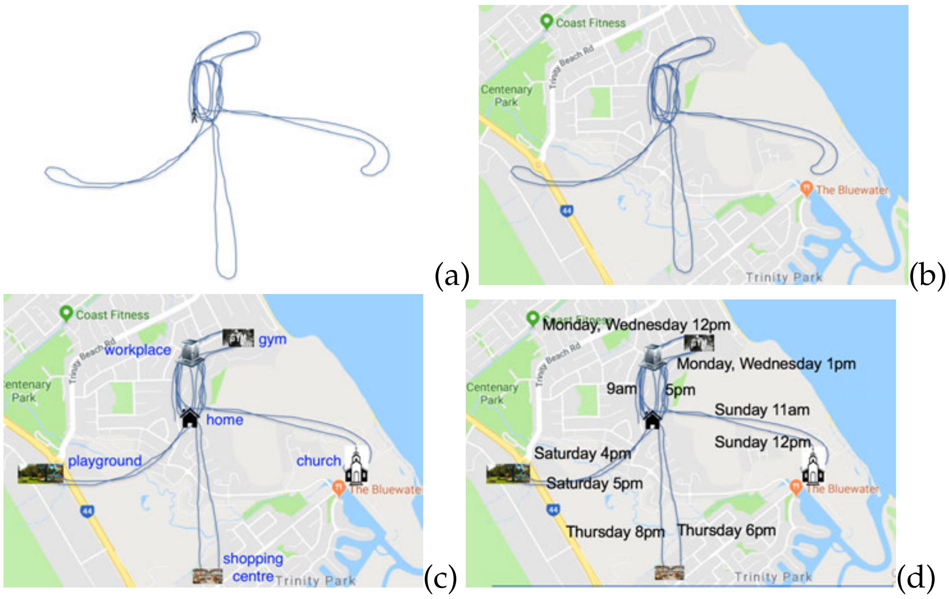


FIGURE 1.7: An example of semantic information.

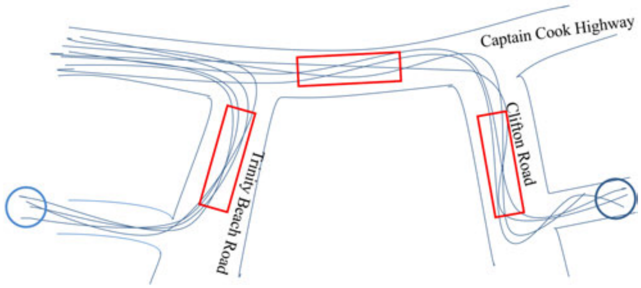


FIGURE 1.8: An example of trajectory paths.

1.3 Research Aim

As previously noted, related work on periodic pattern mining can be categorised into two fields of research: traditional periodic pattern mining and spatio-temporal trajectory periodic pattern mining. The former focuses on traditional datasets, including event/sequence data, time series data and social network data. The latter aims at spatio-temporal trajectory data. This thesis explores this latter research, analyses related studies and develops a new systematic framework to scope six aims, which are:

1. To develop an efficient and effective periodic pattern mining for irregular spatio-temporal trajectories without expensive interpolation process;
2. To develop an efficient and effective periodic pattern mining considering the sequence of trajectory;
3. To develop an efficient and effective periodic pattern mining considering spatiality and temporality at the same time;

4. To develop an efficient and effective periodic pattern mining considering the hierarchy of space ;

5. To develop an efficient and effective periodic pattern mining considering semantic background information;

6. To develop an efficient and effective periodic pattern mining for clustered both trajectory nodes and clustered trajectory paths.

To achieve these six research aims, we propose six hypotheses:

1. A periodic pattern mining for irregularly sampled spatio-temporal trajectories without expensive interpolation process is more efficient than traditional periodic pattern mining approaches whilst capable of detecting effective periodic patterns;

2. Considering the sequence of trajectory is able to avoid unwanted false positive reference spots, and also false negative reference spots;

3. Considering spatiality and temporality at the same time in periodic pattern mining will improve the effectiveness of periodic patterns from spatio-temporal trajectories;

4. Considering the hierarchical nature of space in periodic pattern mining will improve the effectiveness of patterns, and also avoid false positive and false negative reference spots;

5. Considering background semantic information in periodic pattern mining will generate more interpretable and effective periodic patterns;

6. Finding periodic patterns for clustered trajectory paths reveals interesting periodic patterns that cannot be detected by traditional trajectory node focused periodic pattern mining.

Based on these six hypotheses, there are six questions that need to be answered:

1. how can we develop an efficient and effective periodic pattern mining for irregular spatio-temporal trajectories without an expensive interpolation process?

2. how can we develop an efficient and effective periodic pattern mining considering the sequence of trajectory?

3. how can we develop an efficient and effective periodic pattern mining considering spatiality and temporality at the same time?

4. how can we develop an efficient and effective periodic pattern mining considering the hierarchy of space?

5. how can we develop an efficient and effective periodic pattern mining considering semantic background information?

6. how can we develop an efficient and effective periodic pattern mining for clustered trajectory paths along with trajectory nodes?

Figure 1.9 briefly illustrates the overall procedure of the thesis. More detailed information is described in Chapter 3. This research mainly focuses on periodic pattern mining from spatio-temporal trajectories. There are two approaches, one is a path-based approach, the other is a node-based approach. The former focuses on finding periodic patterns for paths, the latter pays attention to obtaining periodic patterns for semantic reference spots. Both approaches also consider the hierarchy of space to find periodic patterns for hierarchical paths and hierarchical semantic reference spots. In addition, in the presence of background semantic information, (human' behaviours are usually related to specific locations) we need to attach semantic information database to this research to consider aspatial background semantic information in periodic pattern mining. We will interpret this specifically in Chapter 3.

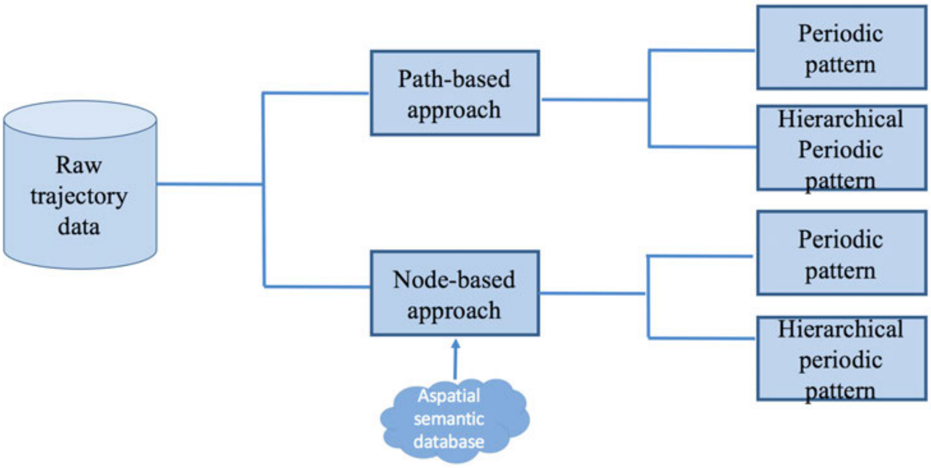


FIGURE 1.9: Overall research of this thesis.

1.4 Contributions

This research makes contributions to the literature of periodic pattern mining to analyse an individual moving object's regular and repeating behaviors based on GPS data. More importantly, this research takes aspatial semantic information and the hierarchy of space into account for periodic pattern mining. The benefit is that more hidden personal movement patterns and behaviours can be inferred. Table 1.1 shows a summary of contributions in this thesis.

TABLE 1.1: Contributions.

Chapter	Publications	Contributions
2	Dongzhi Zhang, Kyungmi Lee and Ickjai Lee (2015), "Periodic Pattern Mining for Spatiotemporal trajectories: A Survey", Proceedings of the 2015 International Conference on Intelligent System and Knowledge Engineering, pp. 306-313.	• to propose a survey for the breath and depth analysis of existing spatio-temporal trajectory periodic pattern mining; • to identify a list of unique characteristics of spatio-temporal trajectories;
4	Dongzhi Zhang, Kyungmi Lee and Ickjai Lee (2018), "Hierarchical Trajectory Clustering for Spatio-temporal Periodic Pattern Mining", Expert Systems with Applications, 92: 1-11.	• to propose a path-based trajectory clustering that takes into account the sequence of trajectory and additional spatio-temporal semantic information in order to produce context-sensitive reference spots (trajectory paths); • to detect hierarchical reference spots (trajectory paths) in order to reduce false positive reference spots (trajectory paths);
5	Dongzhi Zhang, Kyungmi Lee and Ickjai Lee (2018), "Semantic periodic pattern mining from spatio-temporal trajectories", Information Sciences (submitted) Dongzhi Zhang, Kyungmi Lee and Ickjai Lee (2018), "Mining Hierarchical semantic periodic patterns from GPS-collected spatio-temporal trajectories", Expert Systems with Applications (Under revision)	• to propose a novel trajectory representation method to describe a spatio-temporal trajectory as a sequence of semantic episodes that match background aspatial semantic information; • to discover spatially, temporally and semantically aggregated concentrations as reference spots for periodic pattern mining; • to detect regular time periods for each spatio-temporal concentration from irregular trajectories; • to consider all spatiality, temporality, semantics, and hierarchy together in detecting reference spots.
6	Dongzhi Zhang, Kyungmi Lee and Ickjai Lee (2018), "Mining Medical Periodic Patterns form spatio-temporal Trajectories", Proceedings of the 7th International Conference on Health Information Science, Springer, Lecture Notes in Computer Science. Accepted and in press Dongzhi Zhang, Kyungmi Lee and Ickjai Lee (2018), "Multi-level Medical Periodic Patterns from Human Movement Behaviors", Journal of Health Information Science and Systems (Submitted)	• to propose a medical periodic pattern mining framework from spatio-temporal trajectories; • to utilise cutting-edge spatio-temporal periodic pattern mining to identify a set of trajectories (possibly patients and health professionals) exhibiting periodic visits to medical centres; • to find medical periodic patterns from spatio-temporal trajectories; • to propose a multi-level (hierarchical) medical periodic pattern mining framework from spatio-temporal trajectories; • to utilise cutting-edge spatio-temporal multi-level periodic pattern mining to identify a set of trajectories (possibly patients and health professionals) exhibiting periodic visits to medical centres; • to find single-level and multi-level medical periodic patterns from spatio-temporal trajectories;

The innovation of this research includes:①we first detect sequentially, spatially, temporally, semantically and hierarchically aggregated concentrations from trajectories with irregular time intervals using aspatial semantic information to find periodic patterns to reveal people's behaviors based on two effective approaches: path-based approach and node-based approach, ②According to moving object's periodic behaviors, it is possible to infer more hidden personal movement patterns, behaviors and information. For instance, a moving object has a regular movement between two universities, thus we can infer that he/she is a student or teacher.

1.5 Structure of Thesis

- Chapter 2 analyses related studies including traditional periodic pattern mining and spatio-temporal data mining. Traditional periodic pattern mining approaches mainly focus on event/sequence data, time series data and social network data. Spatio-temporal data mining mainly includes association rule mining, sequential pattern mining and periodic pattern mining. We analyse these literatures and identify several research gaps in existing studies in periodic pattern mining.

- Chapter 3 describes the overall framework of this thesis. This chapter briefly interprets the overall procedure, significant definitions and function of each component in this framework.

- Chapter 4 proposes a new path-based trajectory clustering algorithm called Traclus (ST) that takes both the sequence of trajectory and the semantic spatio-temporal information such as direction, speed, and time into account. This algorithm is comprised of two phases: partitioning and grouping. In the first phase, it partitions a trajectory into a few line segments through characteristic points at which the moving object makes a sharp turn. In the second phase, it extends traditional density-based points clustering DBSCAN to group these line segments. In order to overcome the non-hierarchical nature of DBSCAN, this approach utilises HDBSCAN, which is a hierarchical version of DBSCAN capable of handling clusters with different densities, while preserving the scalability of DBSCAN. HDBSCAN is able to generate hierarchical clusters using the single-linkage and then automatically extracts clusters from a hierarchical tree. In single-level, the results indicate our approaches can generate more reference spots (trajectory paths) and periodic patterns than existing approaches. In multi-level, more hidden and meaningful periodic patterns can be found with hierarchical reference spots (trajectory paths).

- Chapter 5 demonstrates the study of hierarchical semantic periodic pattern mining from spatio-temporal trajectories. This node-based approach utilises semantic information extracted from background maps, such as a restaurant, a university and a gym to identify spatially and temporally aggregated dense regions from irregularly sampled trajectories, and then applies Hidden Markov Model (HMM) to identify semantically meaningful stops (places where an object or a user stays more than a user-specified threshold, thereby indicating the object is engaged in a meaningful behaviour). Next, it applies Lomb-Scargle periodogram to find periods for each semantically meaningful stop, and finally mine periodic patterns for each stop-point (previous methods employed Fourier transform for period detection). One limitation of Fourier transform is the use of evenly sampled points as input. It is for this reason that previous approaches use interpolation to make input trajectory regular, or assume that the trajectory is with regular time intervals. In addition, the hierarchy of space is also considered in the process of periodic pattern mining. Experimental results show our method can obtain periodic patterns with higher effectiveness and efficiency than previous approaches, and it is possible to infer more important and meaningful information through the moving object's behaviours and semantic information. As with Chapter 4, more hidden and meaningful periodic patterns can be extracted based on multi-level reference spots when compared to single-level counterparts.

- Chapter 6 is is a case study, which shows periodic pattern mining from spatio-temporal trajectories when applied to a health context.
- Chapter 7 summaries the content and contributions in this thesis, and briefly describes the potential future work that can be further explored based on this thesis.

1.6 Framework

This section proposes a new framework for PPM from spatio-temporal trajectories. Movement trajectories include valuable and repeating information and patterns (Gudmundsson et al., 2017; Han et al., 2010). Periodic patterns can reveal this useful and important information. However, existing studies are not sufficiently adequate to mine periodic patterns using real spatio-temporal trajectory data. This thesis proposes a new framework to obtain more effective and efficient periodic patterns from spatio-temporal trajectories.

Figure 1.10 illustrates the overall framework proposed in this thesis. Our framework is mainly divided into two approaches, a path-based approach and a node-based approach. Both approaches use raw trajectory data as input, and then the path-based approach considers all features of trajectory except irregularity and aspatial semantics to find usual periodic patterns and hierarchical periodic patterns. The node-based approach considers all features of trajectory except path to find periodic patterns and hierarchical periodic patterns. Both approaches are applied to overcome the drawbacks of existing studies but focus on different results. One is for periodic patterns among trajectory paths while the other focuses on periodic patterns among semantic reference spots. As a result, more effective periodic patterns are obtained, based on trajectory paths and semantic reference spots. In addition, a semantic database is attached to the node-based approach to mine semantic periodic patterns for human behaviors.

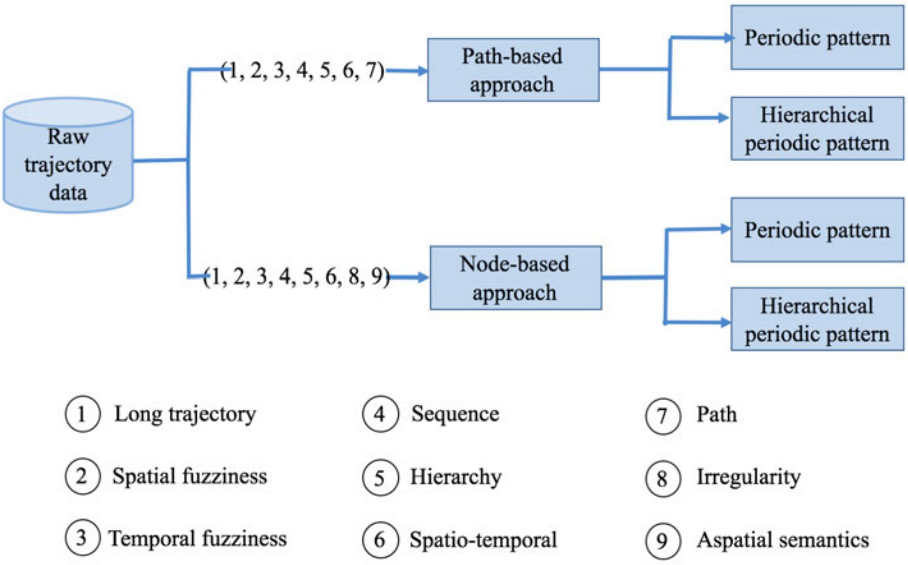


FIGURE 1.10: Overall framework of our proposed method.

According to the literature, there are nine key factors that need to be considered in this research as shown in Table 2.5. Previous studies have covered three items: (① long trajectory; ② spatial fuzziness; ③ temporal fuzziness). Our framework is also to consider further six characteristics: ④ sequence; ⑤ hierarchy; ⑥ spatio-temporal; ⑦ path; ⑧ irregularity; ⑨ aspatial semantics.

Chapter 2

Literature Review

This chapter investigates previous studies for PPM. Section 2.1 briefly provides definitions of the different categories of periodic patterns. Section 2.2 introduces existing studies for traditional PPM, including PPM in event/sequence data, time series data and social network data. In Section 2.3, we describe related works in spatio-temporal trajectory data mining, which mainly involve ARM, sequential pattern mining and in particular, discuss PPM. Section 2.4 summaries the depth review of literature which is related to our research, and proposes objectives and tasks of our research.

2.1 Periodic Pattern

A periodic pattern is a pattern that is defined as repeating behaviors at a certain location (spatial property) with regular time interval (time property) for objects. Three types of periodic patterns can be detected (Sirisha et al., 2013): (1) partial *vs.* full; (2) perfect *vs.* imperfect; (3) synchronous *vs.* asynchronous.

2.1.1 Partial *vs.* Full

A full periodic pattern is a pattern where every element in every position in the pattern exhibits the periodicity, such as, the sequence:

$$ABABBCAB.$$

AB is a full periodic pattern with a period 2. Partial periodic pattern is a pattern where one or more elements do not present the periodicity, for instance, the sequence:

$$ABDABBABD.$$

*AB** is a partial periodic pattern with a period 3 because the third element does not exhibit periodicity. In spatio-temporal trajectory data, a full periodic pattern is ideal but it is more time consuming or impossible with spatio-temporal data due to the uncertainty embedded in trajectories. Therefore, partial PPM is more suited to spatio-temporal trajectories.

2.1.2 Perfect *vs.* Imperfect

A sequence is said to have perfect periodicity, if a pattern *p* with a period *t* starts with the first occurrence of *p* until the end of the sequence, and every next *p* occupies *t* positions away from the current occurrence *p*.

BCEBCFBCA.

*BC** is a perfect periodic pattern with a period 3. It occurs 3 times from the first occurrence to the end of the sequence. Imperfect periodicity means that the pattern deviates from the next expected occurrence.

ABCHDJABD.

Given the above example, *AB** is an imperfect periodic pattern; the pattern is missed in the second expected position. A perfect periodic pattern is also almost impossible (highly unlikely) with spatio-temporal data due to the trajectory uncertainty. Thus, imperfect PPM is more suited to spatio-temporal trajectories.

2.1.3 Synchronous *vs.* Asynchronous

A pattern that occurs periodically without any misalignment or with no intervention of random noise is called a synchronous periodic pattern.

ABCADCBDCBAC.

***C* is a synchronous periodic pattern with a period 3. Asynchronous periodic patterns mean the patterns might be misaligned due to the intervention of random noise. The misalignment is accepted only up to a certain threshold value. In the sequence,

ABCDBCCBABC.

**BC* is an asynchronous periodic pattern due to the insertion of random noise events *CB* between the second and third occurrences of the pattern. Asynchronous PPM can deal with the shift and distortion due to the presence of random noises in the periodic sequence.

In conclusion, for spatio-temporal trajectories, periodic patterns are usually partial, imperfect and asynchronous, due to the mixture of periodic activities and non-periodic activities in real world data.

2.2 Traditional PPM

In existing work, traditional PPM can be divided into three types: one-dimensional event/sequence, two-dimensional time series or spatial data, and social network data.

2.2.1 One-dimensional Event/Sequence PPM

A sequence is an ordered list of elements from any application domain. The order of elements in a sequence might be implied by time order, such as physical positions in DNA, protein sequences, or stock market data (Dong, 2009). If the order of elements is implied by time order, the sequences can be called event sequences.

One-dimensional event PPM is used to find periodic patterns from a set of sequences (symbols or events). Existing techniques for PPM include Apriori (Agrawal and Srikant, 1994) and Max-subpattern Hit Set (Cao et al., 2004). In addition, it is worth noting that period discovery is a key factor for one-dimensional PPM. The more precise period, the more accurate periodic patterns. The main works on period detection are fast Fourier transform and autocorrelation (Kargupta, 2005). Automatic period detection can discover as many periods as possible, but false and redundant periods might occur. The advantage of user-specific periods is that specific periods are obtained for mining periodic patterns, but other periods will be discarded, possibly leading to the loss of some patterns.

To sum up, in Table 2.1, it can be concluded that these discrete symbols are explicitly given with regular and constant time intervals in one-dimensional PPM, but one-dimensional PPM methods cannot be directly applied to two, even three-dimensional data due to the additional information with two/three-dimensional data.

TABLE 2.1: Symbolic PPM for trajectory.

	Space	Time	Automatic Detection of Time Windows
(Agrawal and Srikant, 1994; Cao et al., 2004; Huang and Chang, 2004; Yang et al., 2013)	No	Regular time interval	No
(Kargupta, 2005)	No	Regular time interval	Yes

2.2.2 Two-dimensional PPM

Two-dimensional PPM focuses on mining two-dimensional data that describes anything consisting of two kinds of properties. Similar to one-dimensional PPM, it is assumed that the data is collected at regular and constant time intervals. We will discuss two-dimensional time-series and spatial PPM in this section.

Time Series PPM

Time series data captures the change of data value over time, such as power consumption data in energy companies, stock prices in the financial market, and event logs in computer networks. Research in time series data mining has concentrated on discovering different types of patterns: sequential patterns (Agrawal and Srikant, 1995; Srikant and Agrawal, 1995; Garofalakis et al., 1999), temporal patterns (Bettini et al., 1998), periodic association rules (Özden et al., 1998), partial periodic patterns (Jiawei Han, 1999; Aref et al., 2004; Yang et al., 2002), and surprising patterns (Keogh et al., 2002).

Most early PPM studies belong to this category; one example is shown in Figure 2.1. In general, the data are collected at regular and constant time intervals. X-axis is for time and Y-axis is for values. By connecting contiguous (adjacent) neighbors, we can get a trajectory (Figure 2.1).

Jiawei Han (1999) developed an algorithm called Max-Subpattern Hit Set which creates a max subpattern tree to mine periodic patterns by two scans of the time series database. Aref et al. (2004) extended the Max-Subpattern Hit Set to develop an online incremental version to allow users to modify the threshold and mine periodic patterns in the presence of insertion and deletion updates in the database. Chen et al. (2011) discovered periodic patterns with a given period using the encoded period segments and then applied Max-Subpattern Hit Set to mine periodic patterns. Some variations of periodic patterns are also represented. L. Zhu et al. (2012) worked on mining approximate periodic patterns in

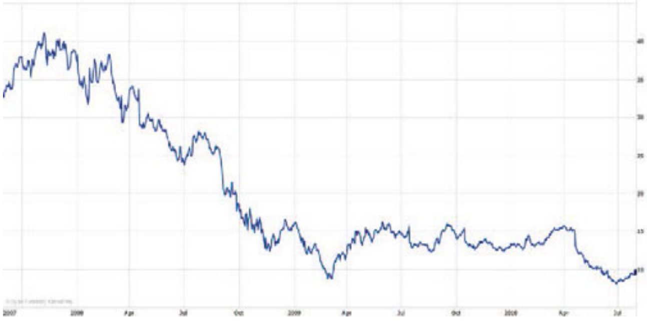


FIGURE 2.1: Change of stock price via regular time intervals (source from: http://static.businessinsider.com/image/4c5862a77f8b9af546280000/image.jpg).

hydrological time series. Another investigation was conducted by Zhang et al. (2007) to find periodic patterns with gap requirements from DNA. Sheng et al. (2006) proposed a method to mine dense periodic patterns in the time series database. Yang et al. (2000), Huang and Chang (2005), Yeh and Lin (2009), and Maqbool et al. (2006) developed algorithms to mine asynchronous periodic patterns. Yang et al. (2001) employed a method to mine surprising periodic patterns. Nishi et al. (2013) proposed an algorithm which uses an apriori-based sequential mining approach to mine flexible periodic pattern. This method has similar drawbacks to apriori-based sequential mining appraoches, such as complexities in mining long sequences. To overcome the drawbacks identified by Nishi et al. (2013), Chanda et al. (2015) proposed an approach to generate flexible periodic patterns which uses a suffix tree to deal with a variable starting position without a mass of redundant computation. Chanda et al. (2017) proposed a weighted PPM algorithm for time series databases.

All existing time series PPM algorithms map two-dimensional data into one-dimensional sequences. Then, one-dimensional PPM is applied to handle the transformed one-dimensional sequences. This is necessary because two-dimensional PPM is designed to handle lower-dimensional data, but not to analyse higher-dimensional data. Therefore, these two-dimensional algorithms cannot be directly applied to three-dimensional spatio-temporal trajectory data; instead, they must be extended/modified to handle additional dimensional information in three-dimensional data in order to not to miss hidden patterns. Table 2.2 shows related studies in time series PPM, and demonstrates these algorithms are not suitable for mining spatio-temporal trajectory data, due to the absence of spatio-temporal properties. L. Zhu et al. (2012), Zhang et al. (2007), Sheng et al. (2006), Yang et al. (2000), Huang and Chang (2005), Yeh and Lin (2009), Maqbool et al. (2006), Yang et al. (2001), Nishi et al. (2013), Chanda et al. (2015), and Chanda et al. (2017) all use automatic period detection algorithms to discover periods.

TABLE 2.2: Time series PPM for trajectory.

	Space	Time	Automatic Detection of Time Windows
(Jiawei Han, 1999; Aref et al., 2004; Chen et al., 2011)	No	Regular time interval	No
(L. Zhu et al., 2012; Zhang et al., 2007; Sheng et al., 2006; Yang et al., 2000; Huang and Chang, 2005; Yeh and Lin, 2009; Maqbool et al., 2006; Yang et al., 2001; Nishi et al., 2013; Chanda et al., 2015; Chanda et al., 2017)	No	Regular time interval	Yes

Spatial PPM

Spatial data involves the data with spatial distributions, such as spatial precipitation patterns, vegetation patterns in selected basins, aquifer properties and change of temperature in different areas (Figure 2.2). Major studies include those by Han et al. (1998) who integrated data cube, bit-mapping and Apriori (Agrawal and Srikant, 1994) to mine segment-wise periodicity with fixed length period. In He et al. (2008), they proposed a multiple partial PPM algorithm in parallel computing environments that detects all valid periods to reduce the cost of communication among processors, while avoiding the generation of redundant patterns. Figure 2.2 shows one example of spatial PPM.

Although spatial properties are considered in spatial PPM, the time element is not embedded in spatial data. The X-axis represents locations, while the Y-axis shows values. Time is not explicit, but a regular or constant time interval is implicitly included (e.g. every second, every minute, every hour, etc.). Table 2.3 shows a summary of past studies in spatial PPM.

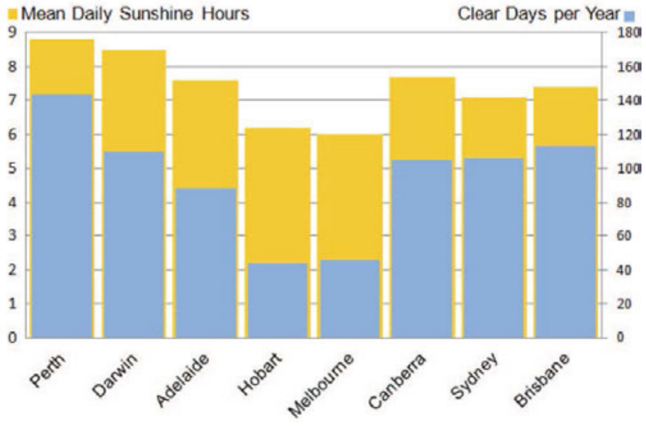


FIGURE 2.2: Sunshine and clear days by capital city in Australia (source from: http://www.aussiemove.com/aus/images/sunshine.png).

TABLE 2.3: Spatial PPM for trajectories.

	Space	Time	Automatic Detection of Time Windows
(Yang et al., 2001)	Yes	Regular time interval	No
(He et al., 2008)	Yes	Regular time interval	Yes

Social Network PPM

PPM in dynamic social networks has been becoming very interesting research nowadays. Analysis of dynamic social networks through PPM was proposed by (Lahiri and Berger-Wolf, 2010; Lahiri and Berger-Wolf, 2008). They proposed a single pass PSEMiner algorithm which uses sub-graphs to capture periodic patterns based on a pattern-tree in polynomial time. Traversal of a pattern tree and creation of many unwanted tree nodes are very time consuming. Apostolico et al. (2011) proposed the ListMiner algorithm to speed up the PSEMiner algorithm by solving unwanted tree node creation. This approach is faster than PSEMiner because it requires traversing of a smaller number of list nodes. However the number of list nodes is still large, and the same graph is stored at different times. This redundant information utilizes substantial memory and time. Halder et al. (2013) proposed an algorithm called SPBMiner which is faster than both PSEMiner and ListMiner. This method stores all entries only once and than finds common sub patterns only once as well. However, a vast number of redundant periodic information is generated, which results in high memory consumption. Halder et al. (2017) proposed a sub graph-based algorithm called SPP miner for PPM. It can also be used for polynomial graph mining. The advantage of this method is that the memory consumption is more efficient compared to other algorithms.

2.2.3 Three-dimensional PPM

Three-dimensional PPM focuses on mining spatio-temporal trajectory data. Obviously, one-dimensional and two-dimensional PPM approaches cannot be used directly for spatio-temporal trajectory PPM. This will be discussed in Section 2.3.3.

2.3 Spatio-temporal Trajectory Data Mining

Three types of spatio-temporal trajectory data mining techniques show relationships between trajectories: Association Rule Mining (ARM); Sequential Pattern Mining (SPM); and PPM.

2.3.1 Spatio-temporal Trajectory ARM

Association rules were first used for supermarket basket data. ARM seeks to discover positive frequent associations among transactions encoded within a database (Agrawal et al., 1993). The aim of ARM is to identify all the strong association rules among itemsets in a given database. The support and confidence of strong association rules must satisfy user-specified minimum support and minimum confidence. ARM finds associations, but fails to find causal effects, sequential patterns and periodic patterns. ARM can be considered for mining trajectory data using Apriori (Agrawal and Srikant, 1994) and FP-tree (Han et al., 2004). For more details about these algorithms, please refer to (Tanbeer et al., 2009; Lee et al., 2009). In addition, Association rules can also be used for location prediction (Verhein and Chawla, 2006; Tao et al., 2004; Yavas et al., 2005).

ARM from spatio-temporal trajectories demonstrates how objects move between places over time (Verhein and Chawla, 2006). For instance, a group of people moves from place p_i to place p_j over time period $[t_i, t_j]$ whilst these moving objects satisfy conditions q, this can be presented as the following:

$$(p_i, t_i, q) \rightarrow (p_j, t_j)[\text{s\%, c\%}],$$

where s and c are support and confidence of the rule, respectively. The support is the number of moving objects that move from place p_i to place p_j during time period $[t_i, t_j]$.

2.3.2 Spatio-Temporal Trajectory SPM

SPM was first proposed by Garofalakis et al. (1999), and since then there have been many studies to detect sequential patterns in time series data (Agrawal and Srikant, 1995; Srikant and Agrawal, 1995; Garofalakis et al., 1999; Bermingham and Lee, 2014). SPM is to detect sequential patterns for a given sequence database satisfying the minimum support threshold (Han et al., 2005). The major algorithms include Generalized Sequential Patterns (GSP) (Srikant and Agrawal, 1995), SPADE (Zaki, 2001), and PrefixSpan (Pei et al., 2001). PrefixSpan is reported to be better than GSP and SPADE (Pei et al., 2004).

Traditional sequential pattern discovery techniques are not readily applicable to SPM in spatio-temporal data. SPM in spatio-temporal trajectory data is more complicated than traditional SPM due to the mixture of temporal and spatial relations. Spatio-temporal SPM for trajectory data has also attracted some attention (Hwang et al., 2005; Liu et al., 2007; Giannotti et al., 2007).

SPM from spatio-temporal trajectories is to find popular transitions from one instance to another which show sequences of geographical locations and regions that a group of moving objects visited in the same order (Agrawal and Srikant, 1995). For instance, a group of people moves from location a to b and then to c: $a \to b \to c$ is a frequent sequential pattern. The typical method is from Giannotti et al. (2007) who propose trajectory pattern mining to enrich the sequential patterns with transition time information between two locations. For instance, a trajectory can be represented as a pair (S, T), where $S = (x_0, y_0), (x_1, y_1), ..., (x_n, y_n)$ is a sequence of points in $\boldsymbol{R}^2$, and $T = (\Delta T_1, \Delta T_2, ..., \Delta T_n) \in \boldsymbol{R}^+$ is a corresponding transition time between two locations. Thus, the trajectory pattern could be presented as:

$$(S, T) = (x_0, y_0) \xrightarrow{\Delta T_1} (x_1, y_1) \xrightarrow{\Delta T_2} (x_2, y_2).$$

The trajectory pattern mining focuses on finding all frequent trajectory patterns (S, T) which satisfy the following condition:

$$support(S, T) \geq sup_{min},$$

where $support(S,T)$ is the number of input trajectories which contains the trajectory pattern (S,T) and sup_{min} is a user-specific minimum support threshold. SPM can find associative effects and sequential patterns, but fails to find periodicity in spatio-temporal trajectories.

2.3.3 Spatio-temporal Trajectory PPM

Spatio-temporal trajectory PPM is attracting increasing attention. Periodicity is a kind of movement rule that naturally exists in moving objects. Moving objects always obey more or less the same route (spatial property) over regular time intervals (temporal property). A periodic pattern (the repeating green line) is shown in Figure 2.3 where an entity exhibits the same spatio-temporal pattern with some periodicity. For example, birds have yearly migration patterns, people travel on regular routes to work and commercial airliners operate regular schedules from one place to another. PPM can be used to discover the intrinsic behavior of moving objects (Han et al., 2010), compressing movement data (Agrawal and Srikant, 1995; Mamoulis et al., 2004), predicting future movements of objects (Han et al., 2010; Jeung et al., 2008), and detecting abnormal events (Han et al., 2010).

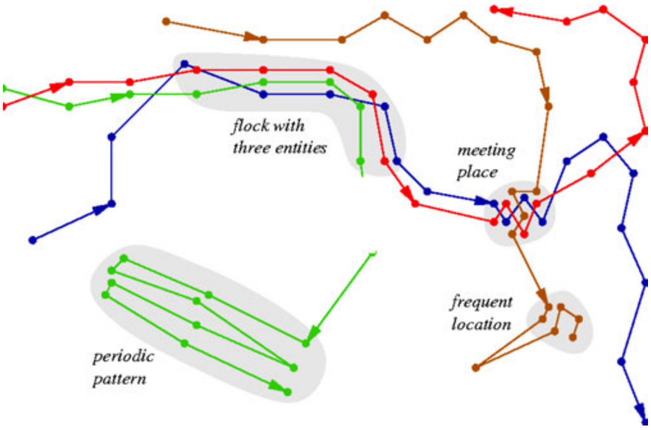


FIGURE 2.3: An example of periodic pattern (Gudmundsson et al., 2017).

Section 2.2 discusses PPM with lower-dimensional data and discusses problems when lower-dimensional PPM is used for spatio-temporal data. Section 2.3.1 and 2.3.2 survey trajectory data mining for ARM and SPM which are not able to detect periodic patterns. These two sections highlight the need for spatio-temporal PPM, which is the main aim of this survey.

2.3.4 Difference among ARM, SPM and PPM

Table 2.4 summarizes the main differences among PPM, ARM and SPM. ARM and SPM typically require a large number of trajectories where the length of trajectory is not important. Conversely, PPM requires a single and extremely long trajectory where the number of trajectories is not important.

PPM has periodicity where the period can be user-specific or automatically detected. However, ARM and SPM do not need to consider periodicity. ARM is not concerned about order, whereas SPM and PPM must strictly obey the time sequence.

TABLE 2.4: Comparison ARM, SPM and PPM.

	Length of Trajectory	Trajectory	Periodicity	Time-Order
ARM	Short	Multiple	No	No
SPM	Short	Multiple	No	Yes
PPM	Long	Single	Yes	Yes

2.3.5 Features of PPM for Spatio-Temporal Trajectories

In the process of PPM for spatio-temporal trajectory data, a number of unique characteristics need to be considered:

- **The Length of Trajectory;**
  ARM and SPM focus on the number of short trajectories. In contrast, PPM uses an extremely long trajectory that might be a month trajectory, one year trajectory or even longer. For instance, Figure 2.4 shows ARM and SPM that focus on a small number of short trajectories. In comparison, In Figure 2.5, PPM focuses on a very long and single trajectory. Thus, mining this extremely long trajectory needs a different approach. PPM finds repetitive patterns from this long single trajectory, whereas it is not suited for use with multiple trajectories.

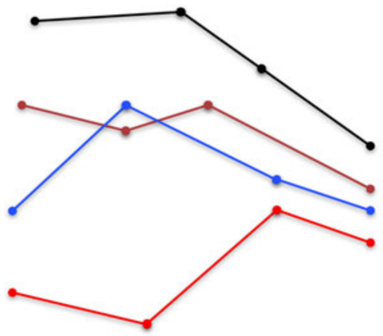


FIGURE 2.4: An example of a group of short trajectories.

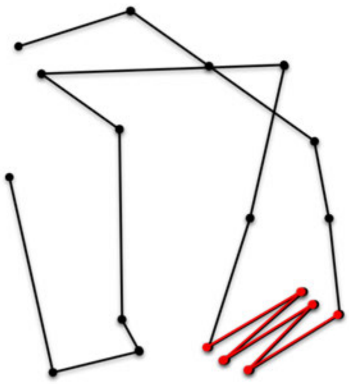
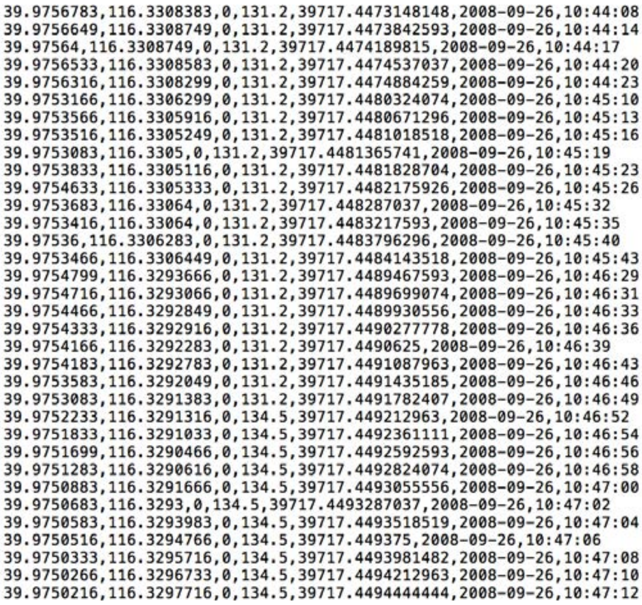


FIGURE 2.5: An example of a long and single trajectory.

- **Irregular Time Intervals;**
  Unlike one- and two-dimensional data, spatio-temporal data is collected at different time intervals due to different sampling rates as shown in Figure 2.6. In the last column, the time intervals are irregular among trajectory nodes. Dealing with this irregularity is a challenging task, but worth exploring.



```
39.9756783,116.3308383,0,131.2,39717.4473148148,2008-09-26,10:44:08
39.9756649,116.3308749,0,131.2,39717.4473842593,2008-09-26,10:44:14
39.97564,116.3308749,0,131.2,39717.4474189815,2008-09-26,10:44:17
39.9756533,116.3308583,0,131.2,39717.4474537037,2008-09-26,10:44:20
39.9756316,116.3308299,0,131.2,39717.4474884259,2008-09-26,10:44:23
39.9753166,116.3306299,0,131.2,39717.4480324074,2008-09-26,10:45:10
39.9753566,116.3305916,0,131.2,39717.4480671296,2008-09-26,10:45:13
39.9753516,116.3305249,0,131.2,39717.4481018518,2008-09-26,10:45:16
39.9753083,116.3305,0,131.2,39717.4481365741,2008-09-26,10:45:19
39.9753833,116.3305116,0,131.2,39717.4481828704,2008-09-26,10:45:23
39.9754633,116.3305333,0,131.2,39717.4482175926,2008-09-26,10:45:26
39.9753683,116.33064,0,131.2,39717.448287037,2008-09-26,10:45:32
39.9753416,116.33064,0,131.2,39717.4483217593,2008-09-26,10:45:35
39.97536,116.3306283,0,131.2,39717.4483796296,2008-09-26,10:45:40
39.9753466,116.3306449,0,131.2,39717.4484143518,2008-09-26,10:45:43
39.9754799,116.3293666,0,131.2,39717.4489467593,2008-09-26,10:46:29
39.9754716,116.3293066,0,131.2,39717.4489699074,2008-09-26,10:46:31
39.9754466,116.3292849,0,131.2,39717.4489930556,2008-09-26,10:46:33
39.9754333,116.3292916,0,131.2,39717.4490277778,2008-09-26,10:46:36
39.9754166,116.3292283,0,131.2,39717.4490625,2008-09-26,10:46:39
39.9754183,116.3292783,0,131.2,39717.4491087963,2008-09-26,10:46:43
39.9753583,116.3292049,0,131.2,39717.4491435185,2008-09-26,10:46:46
39.9753083,116.3291383,0,131.2,39717.4491782407,2008-09-26,10:46:49
39.9752233,116.3291316,0,134.5,39717.449212963,2008-09-26,10:46:52
39.9751833,116.3291033,0,134.5,39717.4492361111,2008-09-26,10:46:54
39.9751699,116.3290466,0,134.5,39717.4492592593,2008-09-26,10:46:56
39.9751283,116.3290616,0,134.5,39717.4492824074,2008-09-26,10:46:58
39.9750883,116.3291666,0,134.5,39717.4493055556,2008-09-26,10:47:00
39.9750683,116.3293,0,134.5,39717.4493287037,2008-09-26,10:47:02
39.9750583,116.3293983,0,134.5,39717.4493518519,2008-09-26,10:47:04
39.9750516,116.3294766,0,134.5,39717.449375,2008-09-26,10:47:06
39.9750333,116.3295716,0,134.5,39717.4493981482,2008-09-26,10:47:08
39.9750266,116.3296733,0,134.5,39717.4494212963,2008-09-26,10:47:10
39.9750216,116.3297716,0,134.5,39717.4494444444,2008-09-26,10:47:12
```

FIGURE 2.6: An example of GPS data (source from: research.microsoft.com).

- **Densely Recorded or Sparsely Recorded;**
  The sampling rate has a great influence on the accuracy of mining. For instance, the presence of sparse data due to a low sampling rate requires trajectory interpolation as depicted in Figure 2.7(a). On the other hand, the presence of dense data due to a high sampling rate requires trajectory smoothing, while ensuring minimum loss of

information. Figure 2.7(b) shows a simplification method, Douglas-Peucker algorithm (Douglas, 1973).

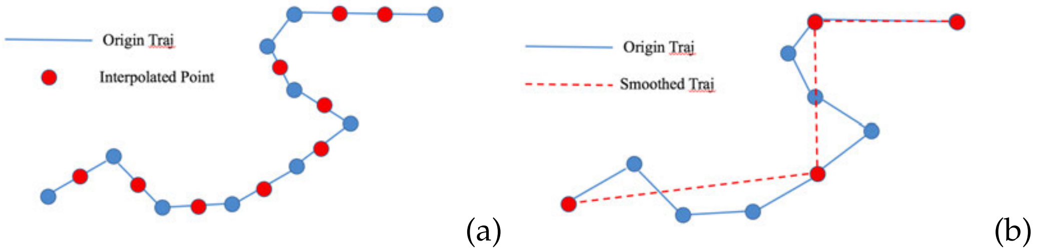


FIGURE 2.7: Trajectory interpolation and smoothing (Douglas (1973)).

- **Trajectory Uncertainty;**
  While we live in a continuous space, spatio-temporal trajectories are recorded at discrete locations. However, imperfect sensors, coarse resolutions in GPS or GSM (a few meters (GPS) and kilometers GSM) (Berberidis et al., 2002), errors in positioning devices, software malfunction or human error lead to the generation of missing or noisy data. For instance, Figure 2.8 shows a spatio-temporal trajectory has three missing data and one noisy data. Thus, it is necessary to include an adequate data preprocessing method to deal with noisy raw trajectory data.

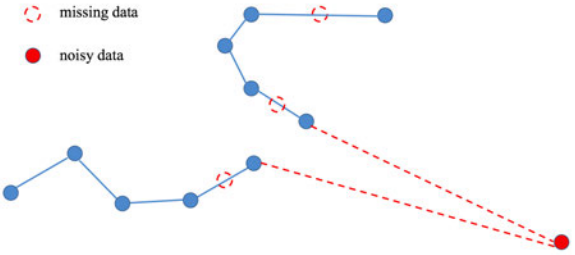


FIGURE 2.8: An example of trajectory uncertainty.

- **Locational (spatial) Fuzziness;**
  Because of spatial uncertainty, traditional PPM cannot be directly applied to trajectory data since the repetitions of spatial locations (x, y) might not exactly coincide. Although objects move along the same route regularly, they might not appear at the exactly the same location. For example, Figure 2.9 shows a person may go from home to office along the same route every day between 9.00 A.M. and 10.00 A.M. However, it is unlikely that he will be at exactly the same location (x, y) on his route every day at 9.30 A.M. (Cao et al., 2007). Cao et al. (2007), Mamoulis et al. (2004), Li et al. (2010a), Jindal et al. (2013), Li et al. (2012), and Li et al. (2011) use clustering approaches to solve uncertainties of spatial locations.

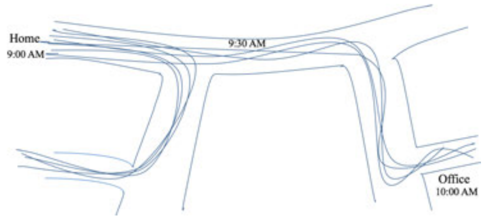


FIGURE 2.9: An example of spatial fuzziness.

- **Temporal Fuzziness;**
  Periodicity is complex since it is hierarchical including partial time span and multiple interleaving periods. The former occurs when periods exist in the partial time interval but not the whole time interval. The latter happens when many periods are interleaved. For example, from March to September; there might be three periods (day, week and month) that interleave each other. For instance, Figure 2.10 shows Bob's daily life. He goes to work from home on weekdays, goes to a restaurant from home on Saturday and goes to a shopping mall on Sunday. The periods include 24 hours for work and 168 hours (7 days) for dinner and shopping. Each period of specific activity is considered a partial time interval. Li et al. (2010a), Jindal et al. (2013), Li et al. (2012), and Li et al. (2011) employ automatic period detection approaches to be capable of finding partial time span and multiple interleaving periods.

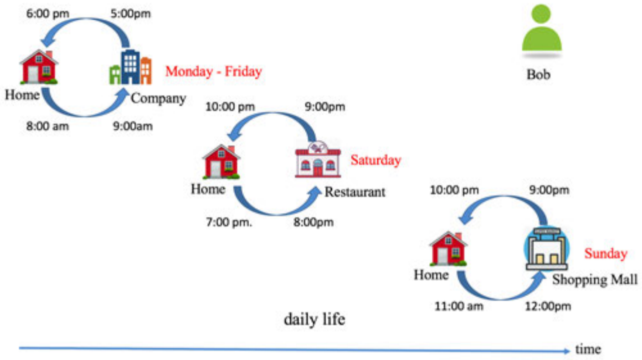


FIGURE 2.10: An example of temporal fuzziness.

- **Hierarchical Spatial Clusters;**
  Spatial data is hierarchical in nature. Spatial characteristics of data can be expressed in different levels of detail. The notion of hierarchical spatial clusters supports hierarchical partitions of the space, by applying hierarchical design to the trajectory data. The design of a hierarchical method to explore and discover different levels of dense regions is a domain-specific task and worthy of exploration. For example, the space can be divided into multiple locations, such as cities, states and countries. As we mentioned in Chapter 1, Figure 2.11

shows different levels of density within regions, such as Queensland as one of the states in Australia, Cairns as one of the cities in Queensland and Trinity Beach as one of the suburbs of Cairns.

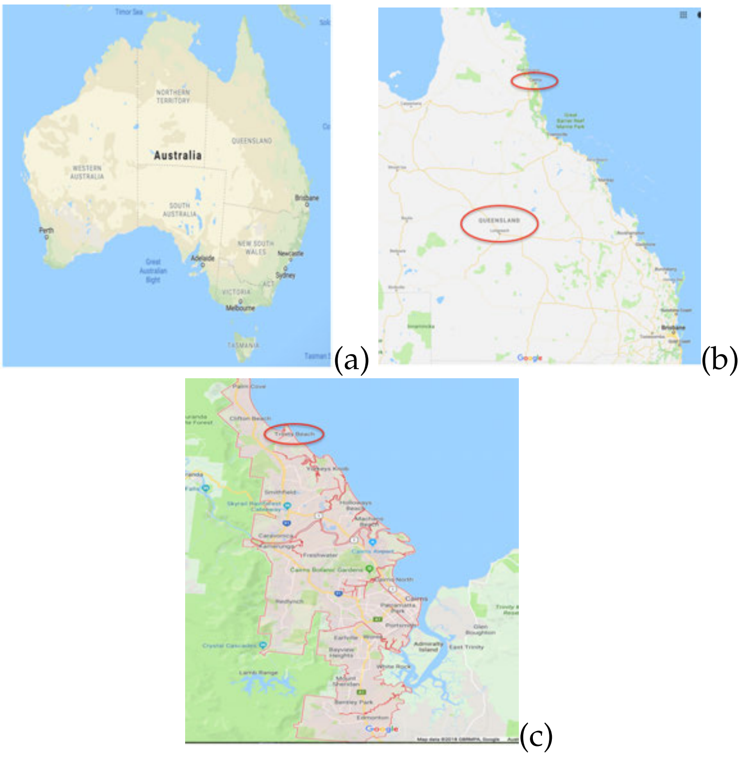


FIGURE 2.11: An example of hierarchy of space.

- **Spatio-Temporal Aspects;**
  Spatial and temporal properties are two core features of spatio-temporal trajectory data. These core features must be considered at the same time in spatio-temporal data mining. For instance, Figure 2.12 shows a real GPS trajectory from home to university. Each trajectory node is a triple data which includes a geographical location and a corresponding timestamp, they are indiscerptible. As an example, the trajectory node that is close to home (145.686, -16.818, "2017-07-30, 06-23-40"), (145.686, -16.818) is geographical location whilst "2017-07-30, 06-23-40" is timestamp.

- **Background Semantic Information;**
  Although mobile devices can track trajectory data, these data do not explicitly represent background semantic geographic information which is relevant to application. This complicates trajectory analysis and PPM. Thus, it is an essential step to integrate trajectory with semantic geographical information in order to capture useful and underlying information (Alvares et al., 2007b; Alvares et al., 2007a; Alvares et al., 2007c). By incorporating this semantic background information, we can explain the spatio-temporal trajectory in a more interesting and meaningful way. For instance, we can identify the

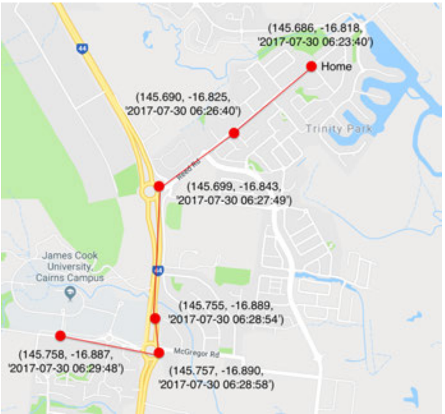


FIGURE 2.12: An example of spatial and temporal aspects.

purpose of the moving object as it travels to different locations. Figure 2.13 illustrates how a moving object starts from home, then goes to his/her workplace, and finally reaches to a supermarket.

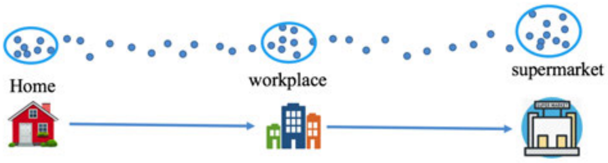


FIGURE 2.13: An example of semantic background information.

- **Sequence of Trajectory;**
  Spatio-temporal trajectory is a sequence comprised of successively sampled points. Each trajectory segment presents two continuous sampled points. Thus, each sampled point is associated with the previous sampled point and the next one, rather than being independent. For instance, Figure 2.14 displays a spatio-temporal trajectory, trajectory nodes 1 and 2 are two end points of a trajectory segment, similarly trajectory nodes 2 and 3 are two end points of another trajectory segment.

- **Trajectory Path;**
  The path which an object follows is called its trajectory, thus path is one important feature of spatio-temporal trajectory. For instance, Figure 2.15 illustrates two trajectory paths: one is from Captain Cook Highway to Clifton Road, the other is Trinity Beach Road coming off Captain Cook Highway. These two trajectory paths are main points of interest in this movement and they represent the main trace of this moving object. In some scenarios, these trajectory paths are main points of interest instead of trajectory nodes (in this example they are just entrances of particular road).

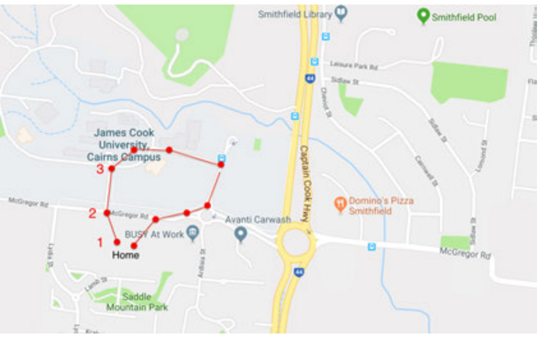


FIGURE 2.14: An example of sequence of trajectory.

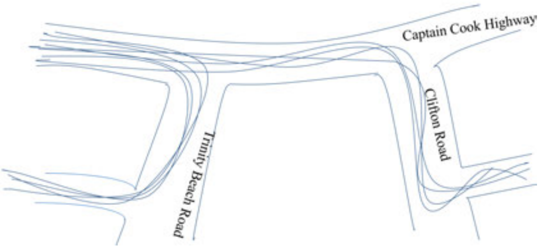


FIGURE 2.15: An example of trajectory path.

2.3.6 Previous Studies in PPM for ST Trajectories

Due to the unique features in spatio-temporal trajectory data, traditional one-dimensional and two-dimensional PPM algorithms cannot be directly applied to mining spatio-temporal trajectory data. Researchers in this field have devised several methods over the past decade to mine spatio-temporal trajectory periodic patterns. Major existing studies can be divided into two groups: the fixed period approach (Cao et al., 2007; Mamoulis et al., 2004) and the reference spot approach (Li et al., 2010a; Jindal et al., 2013; Li et al., 2012; Li et al., 2011).

Fixed Period Approach

This approach proposes an algorithm to find maximal periodic patterns with a user-specific period from a long spatio-temporal data. This algorithm first segments the long trajectory data into sub segments where the length of the sub-segment is the length of the user-specified period. To overcome the drawbacks of regular grid, this approach uses traditional data mining algorithm density-based clustering (DBSCAN (Ester et al., 1996)) to discover dense clusters as valid regions (using cluster ids to replace trajectory data) and hash-based methods to speed up the algorithm. The fixed period approach finds all frequent singular patterns (1-patterns). Finally, it uses bottom-up and top-down mining techniques to generate longer periodic patterns. The main flow for the fixed period approach is shown in Figure 2.16.

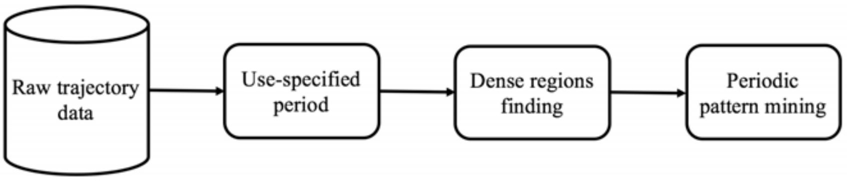


FIGURE 2.16: Overall procedure of the fixed period approach.

First, this approach segments the raw long trajectory into smaller sub-trajectories based on a certain time period (user-specific period). Second, it utilises density-based clustering DBSCAN to find dense regions and distributes class labels to these regions (as shown in Figure 2.17). The approach then transforms the raw trajectory data into one-dimensional event/sequence data. Finally, it uses traditional one-dimensional PPM algorithm Max-Subpattern Hit Set to mine periodic patterns.

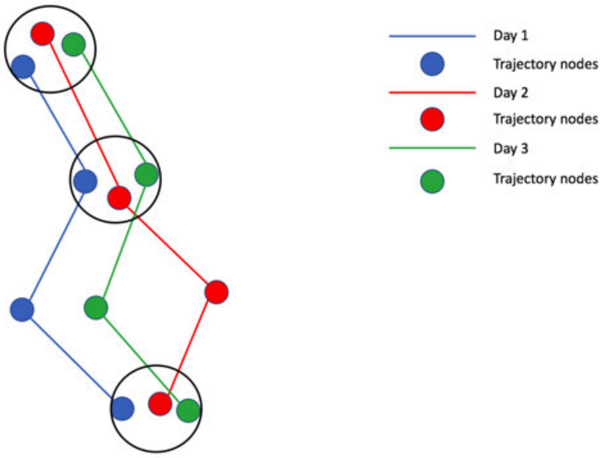


FIGURE 2.17: One long trajectory is divided into 3 segments (red, green and blue). Each segment represents a period with one day, and then a clustering method is used to discover dense regions (regions with circles).

To sum up, the fixed period suffers from several major drawbacks. First, the approach takes a polished and preprocessed(regularly sampled) trajectory as an input. There is no effective data preprocessing method embedded to deal with irregularly sampled raw trajectories. Note that, trajectories could have different sampling rates, and need preprocessing before they are analyzed. Second, this approach assumes that the data is collected at regular and constant time intervals. The approach does not consider this irregularity. Apparently, an irregular time interval is more likely the case in the real world. This irregularity must be resolved using trajectory simplification or interpolation before PPM. Third, this approach uses user-provided periods to mine periodic patterns. Therefore, the approach is user-centered rather data-centered, and leads to a possible loss of periods and patterns. Fourth, temporal information is not considered;

instead, it is abstracted into discrete regional symbols. Thus temporal property is hidden in segment symbols, or is represented by redundant symbols. As a result, hidden and valuable patterns cannot easily be mined. Fifth, hierarchical spatial clusters and hierarchical temporal periods are not considered to mine inherent hierarchical dense regions and time intervals. Sixth, extra computations are required to merge and sort segmented trajectories repeatedly, which leads to extra time and inefficiency of the overall processing performance. Finally,the fixed period approach focuses on trajectory nodes rather than paths.

Reference Spot Approach

The reference spot approach first proposes a two stage algorithm for automatic period detection and periodic behavior mining. In the first stage, this approach uses the kernel method (Worton, 1989) to calculate densities to find reference spots (dense regions) that are frequently visited by moving objects. The approach then obtains periods associated with each reference spot using both Fourier transform and autocorrelation (Kargupta, 2005). For the presence/absence of a time series, a given timestamp value 0 represents when an individual was out of a dense region and value 1 means when an object was in this dense region. Since every period is associated with at least one reference spot, all periods in the movement are guaranteed to be detected when attempting to detect the periods associated with each reference spot. In the second stage, periodic behaviors can be generated by considering all the reference spots associated with a period. The authors extended this work to handle missing data interpolation and movement prediction (Li et al., 2012). The main flow for the reference spot approach is shown in Figure 2.18.

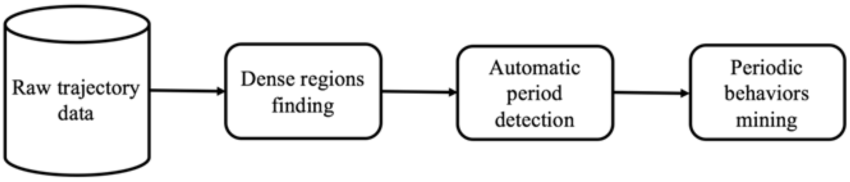


FIGURE 2.18: Overall procedure of the reference spot approach.

The drawbacks of this approach are as follows. First, similar to the fixed period approach, this method does not have an adequate data preprocessing method to process raw inconsistent trajectory data with different time stamps. Second, time intervals are assumed to be regular. That is, data are sampled at regular intervals (for instance every hour); thus, how to deal with irregular time intervals still needs to be explored in the method. Third, this approach only proposes to compare the actual Fourier power spectrum values with the threshold based on randomly permutated presence or absence of event sequence. Thus Markov model of time series is ignored. The probability of a moving object in an area at a timestamp is largely affected by the pervious timestamp. Fourth, the reference spot approach improves the period detection method (Bar-David, 2009) to handle noise and also detect multiple interleaved partial periods. However, it does not find multi-level hierarchical periods and thus could miss some valid periods. Fifth, the temporal property is not considered when this approach discovers reference spots (dense regions), thus this approach does not really take into account both important aspects of spatio-temporal data. Integrating temporal property can obtain more precise estimation of reference spots. Sixth, this approach still does not consider the hierarchical nature of space and time. Finally, a reference spot still focuses on finding periodic patterns based on the trajectory node instead of the trajectory path.

2.4 Summary of Algorithms

In Table 2.5, eight important features are listed with major spatio-temporal PPM methods. The table clearly shows that the two major approaches in the literature are not able to handle all of these important features. For instance, two approaches can handle trajectories with arbitrary length, but they fail to deal with irregular time intervals in the raw trajectory data. They do not consider spatial hierarchies, sequence of trajectory, spatio-temporal aspects together and background semantic information. Our research aims at solving these six issues in PPM. In addition, the existing PPM approaches only focus on finding periodic patterns for the trajectory node instead of the trajectory path, thus, our approach also considers periodic patterns among trajectory paths.

TABLE 2.5: Comparison of spatio-temporal PPM approaches.

	Fixed Period	Reference Spot	Our Proposed Framework
Long Trajectory	✓	✓	✓
Temporal Fuzziness	×	✓	✓
Location Fuzziness	✓	✓	✓
Irregularity	×	×	✓
Spatio-temporal	×	×	✓
Semantics	×	×	✓
Hierarchy	×	×	✓
Sequence	×	×	✓
Path	×	×	✓

Chapter 3

Preliminaries and Definitions

This chapter introduces the concepts and definitions which are related to PPM in this thesis. Section 3.1 briefly discusses the aspatial semantic information database. Section 3.2 provides key terminologies related to PPM from spatio-temporal trajectories in this thesis. Finally, Section 3.3 summary the content of this chapter.

3.1 Aspatial Semantic Database

This research requires the presence of aspatial semantic information in the process of PPM. A geographic information database is the way of extracting aspatial semantic information. This database can be used to annotate the type of places, such as a restaurant or a shopping mall. OpenStreetMap [3] is used in this research to obtain aspatial semantic information. OpenStreetMap is a very large, free geographic database that covers all countries and includes millions of place names.

3.2 Preliminaries and Definitions

The aim of this section is to introduce correlative definition and terms utilized in this research.

- **A spatio-temporal trajectory**: A **spatio-temporal trajectory**, T, is a list of spatio-temporal entries, $(\langle x_1, y_1, t_1\rangle, \langle x_2, y_2, t_2\rangle, \ldots, \langle x_n, y_n, t_n\rangle)$, where $x_i, y_i \in \boldsymbol{R^2}$ and $t_i \in \boldsymbol{R}^+$ for $1 \leq i \leq n$ and $t_1 < t_2 < \ldots < t_n$. A **regular spatio-temporal trajectory** is when $|t_{j+1} - t_j| = |t_{k+1} - t_k|$ for $\forall\ j, k$ where $1 \leq j \neq k < n$ whilst a **irregular spatio-temporal trajectory** is when $|t_{j+1} - t_j| \neq |t_{k+1} - t_k|$ for $\exists\ j, k$ where $1 \leq j \neq k < n$. For instance, Figure 3.1 shows spatio-temporal trajectories, where X-axis and Y-axis represent the location, and Z-axis means time. The green points present trajectory nodes in T, and every node is composed of a triple (x, y, t). In GPS data, x is longitude and y is latitude.

  GPS data: Table 3.1 shows the format of GPS data. GPS data usually include $\langle id, longitude, latitude, altitude, date, time\rangle$. In spatio-termporal trajectory data mining, some of these attributes are usually used,

[3] https://www.openstreetmap.org

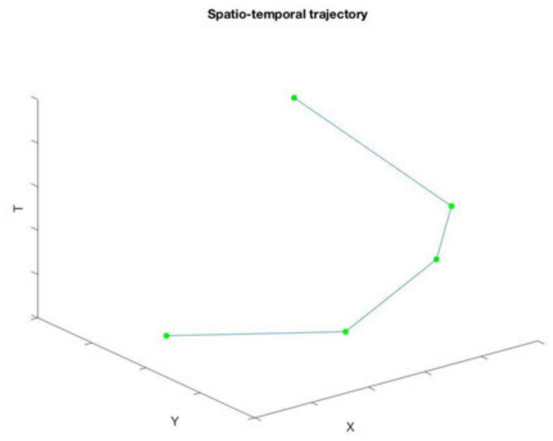


FIGURE 3.1: An example of spatio-temporal trajectory.

including id, longitude, latitude, timestamp (date + time), where id is the identification number of the moving object, latitude and longitude represent spatial geographic coordinates and date and time present the timestamp. PPM focuses on individual moving object, thus we only choose geographic information (longitude and latitude), and time stamp, but do not consider id attributes. In addition, we do not consider altitude in PPM.

TABLE 3.1: An example of GPS trajectory data.

id	longitude	latitude	altitude	date	time
1	39.45892	116.381293	490	2016-01-01	18:01:01

- **A reference spot:** A reference spot (dense region) is a specific spatial area that the moving object frequently visits. Different from ARM and SPM, this research focuses on finding reference spots from individual moving object. In Chapter 4, we also refer to trajectory paths as reference spots. If a reference spot (trajectory path) has a non-zero period, this reference spot (trajectory path) is a periodic path. In Chapter 5, we focus on finding periodic patterns among reference spots in the presence of semantic background information. Therefore we refer to these reference spots as semantic reference spots.

- **A semantic spatio-temporal trajectory:** T_{sem}, is a list of spatio-temporal, semantically annotated entries, $(\langle x_1, y_1, t_1, a_1\rangle, \langle x_2, y_2, t_2, a_2\rangle, \ldots, \langle x_n, y_n, t_n, a_n\rangle)$, where $x_i, y_i \in \boldsymbol{R^2}$, $t_i \in \boldsymbol{R}^+$, and $a_i \in \mathcal{A}$, for $1 \leq i \leq n$ and $t_1 < t_2 < \ldots < t_n$. $\mathcal{A}$ is a finite set of semantic labels which is from a geographical semantic database, such as a restaurant or a university.

- **A semantic episode:** A semantic episode $E_{j,k}$, is a single vector that represents a portion of movement from a semantic trajectory, T_{sem}. The portion of movement starts from the trajectory's j-th index and ends at its k-th index (inclusive), where $j \leq k$. Note that a semantic episode always maximises the number of contiguous entries in T_{sem} with the same semantic label, which means that $T_{sem}.a_j = T_{sem}.a_{j+1} = \ldots = T_{sem}.a_k \wedge T_{sem}.a_{j-1} \neq T_{sem}.a_j \wedge T_{sem}.a_{k+1} \neq T_{sem}.a_k$. The actual vector of the semantic episode $E_{j,k}$ contains a geometry, g, which represents a portion of an entity's movement from indices j to k; a start and end time; t_s and t_e respectively; and a semantic label, a. Specifically, $E_{j,k} = \langle g, t_s, t_e, a \rangle$, where g is constructed using $\{T_{sem}.xy_j, T_{sem}.xy_{j+1}, \ldots, T_{sem}.xy_k\}$, $t_s = T_{sem}.t_j$, $t_e = T_{sem}.t_k$, and $a = T_{sem}.a_j = T_{sem}.a_{j+1} = \ldots = T_{sem}.a_k$.

  The geometry, g, of a semantic episode is often used to infer additional semantic information. For example, the geometry of a semantic episode with the semantic label $\{Stop\}$ (a stop episode), is useful to discover the real-world place where the stop occurred. However, there are many types of semantic episodes though: how g is therefore constructed varies. In this paper, a semantic stop episode means that a user is staying in a particular place more than the user-specified duration doing a meaningful activity.

- **A semantic reference spot:** A semantic reference spot is a spatial place with annotated type. It can be represented as $(RS, type)$, *RS* means reference spot, *type* represents a type of reference spot, such as a restaurant, university or hospital.

- **A hierarchical reference spot:** A hierarchical reference spot represents a division of a reference spot or merger of reference spots at a previous step. In a trajectory path, for instance, Figure 3.2(a) shows a hierarchical trajectory path (in red curve) which includes the paths in Trinity Beach Road, Clifton Road and Captain Cook Highway. In a semantic reference spot, for example, Figure 3.2(b) demonstrates that State A includes City A and City B. City A is composed of many buildings, such as business premises, gyms, restaurants and shopping malls. A shopping mall may have many shops as well. Hospital, supermarket and zoo can be combined as City B. A hierarchical reference spot can be built as a dendrogram, which is a type of tree diagram showing hierarchical reference spots.

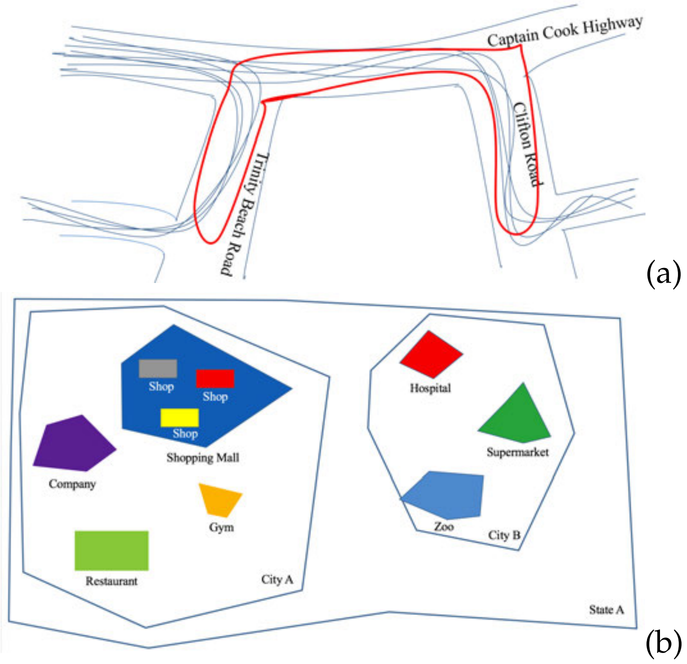


FIGURE 3.2: An example of hierarchy of space.

3.3 Summary

In this chapter, we introduce related concepts which are relative to periodic pattern mining in this thesis. In the later chapters, we will discuss our approaches in detail.

Chapter 4

Hierarchical Trajectory Clustering for Spatio-temporal PPM

In this chapter, we present a study of path-based approach for PPM from spatio-temporal trajectory. In Section 4.1, we present an introduction of this study. Section 4.2 shows an overall framework for extracting (hierarchical) periodic patterns based on trajectory paths. Experimental results are presented and discussed in Chapter 4.3. We summarise our result in Section 4.4.

4.1 Introduction

This chapter aims at developing a path-based PPM approach, and compares these patterns against traditional node-based periodic patterns. In the process of PPM, there are four features of spatio-temporal trajectory data need to be considered, ① sequence of trajectory; ② spatial and temporal together; ③ hierarchy of space; ④ trajectory path.

Figure 4.1 displays an example illustrating the drawbacks of traditional approaches. Let us take an example of 10 trajectory nodes $(p_a, p_b, p_c, \ldots, p_i, p_j)$ where 4 black circles (p_a, p_d, p_g, p_j) for 'home', blue triangles (p_b, p_e, p_h) for 'work' and red squares (p_c, p_f, p_i) for 'gym' as shown in Figure 4.1(a). For better explanation, 'home', 'work' and 'gym' are used as a dense region, and they can also be some trajectory paths. Figure 4.1 (b) and (c) illustrate two possible trajectories: T_A and T_B. T_A exhibits a pattern of moving from 'home' to 'work' and then go to 'gym' before they come back 'home' whilst T_B shows a pattern of moving from 'home' to 'work' and come back 'home' (might be weekdays), and 'home' to 'gym' and come back 'home' (might be weekends). These two trajectories exhibit totally different movement behaviours (sequences), and they have different reference spots and thus have different periodic patterns. However, traditional spatio-temporal PPM approaches (Cao et al., 2007; Li et al., 2012) ignore these sequences but instead only consider the 10 trajectory nodes as a set of unrelated (unordered) targets when detecting reference spots. That is, traditional approaches fail to distinguish the two trajectories T_A and T_B, but detect the same set of two reference spots for them. In addition, traditional approaches cannot separate and detect two interesting places of 'work' and 'gym', but rather detect them as one reference spot. This is due to the

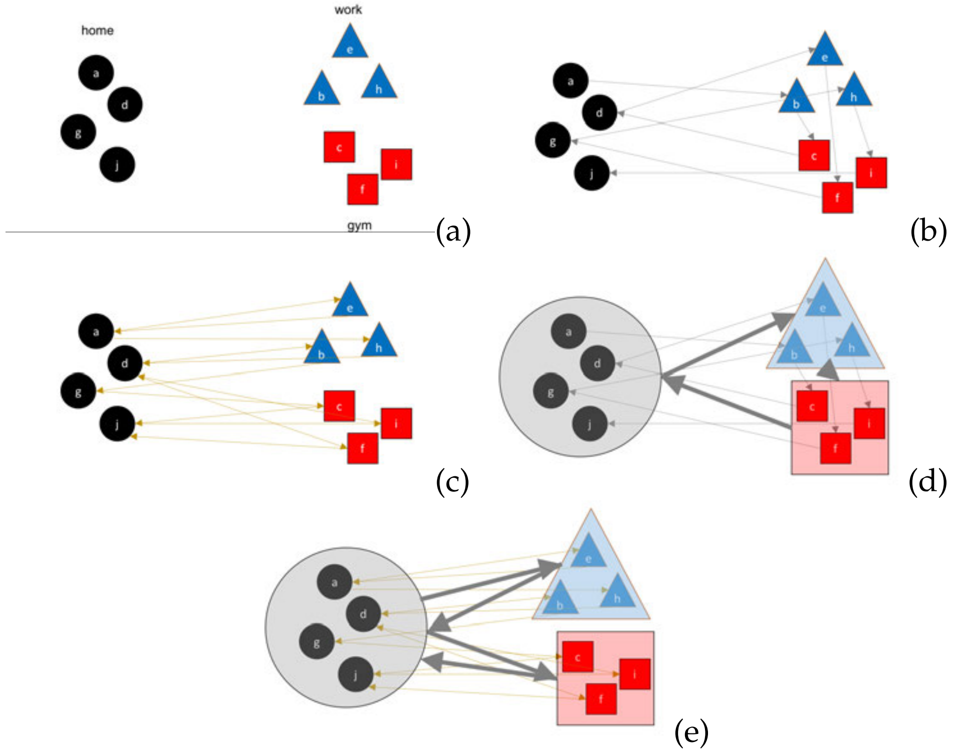


FIGURE 4.1: An example illustrating drawbacks of traditional approaches: (a) 10 points $(p_a, p_b, p_c, \ldots, p_i, p_j)$ where 4 black circles (p_a, p_d, p_g, p_j) for 'home', blue triangles (p_b, p_e, p_h) for 'work' and red squares (p_c, p_f, p_i) for 'gym'; (b) Person A's trajectory $T_A = \{p_a, p_b, p_c, p_d, p_e, p_f, p_g, p_h, p_i\}$; (c) Person B's trajectory $T_B = \{p_a, p_e, p_a, p_h, p_g, p_c, p_j, p_i, p_d, p_f, p_j\}$; (d) Person A's trajectory exhibiting a 'home'-'work'-'gym' sequential pattern; (e) Person B's trajectory exihibiting 'home'-'work' and 'home'-'gym' patterns.

ignorance of trajectory sequence, and also due to the underlying density-based nature of clustering. Figure 4.1(d) shows periodic patterns between 'home' and 'work', 'work' and 'gym', and 'gym' and 'home' whilst Figure 4.1(e) exhibits periodic patterns between 'home' and 'work', and 'home' and 'gym'. There is no direct periodic patterns between 'work' and 'gym' in Figure 4.1(c). One thing to note is that there is a hierarchical cluster 'work-and-gym' combining 'work' and 'gym' in Figure 4.1(d) which could imply a potential hierarchical periodic pattern of T_A between 'home' and 'work-and-gym'.

In Figure 4.2, although the existing node-based approach can find some reference spots around home (trajectory nodes $\{a, d, g, j\}$, work $\{b, e, h\}$ and gym $\{c, f, i\}$, it fails to find some trajectory paths, such as $p(x, z)$, $p(t, s)$, $p(n, w)$. Sometimes, finding periodic patterns among trajectory paths is very important and useful according to the specific applications. For instance, a route planner can generate strategies along the paths that a moving object frequently and regularly passes to provide input to road planning or business premises.

In this chapter, we start to identify drawbacks in the process of

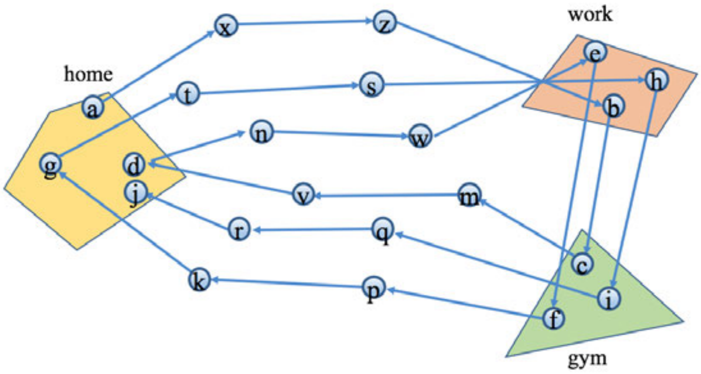


FIGURE 4.2: Another example illustrating the drawbacks of traditional approaches.

discovering reference spots for PPM and utilises our algorithm to reflect the sequence of objects for finding reference spots (trajectory clustering method which considers the sequence of objects and more attributes). Then we extend it to hierarchical clustering to incorporate the hierarchy of objects. We compare and contrast our algorithm with node-based clustering such as Kernel function, Grid-based and initial Traclus to highlight the importance of considering the hierarchy, sequence and spatio-temporal aspects. Note that, in this chapter, we refer to trajectory paths as reference spots in order to make comparisons with the existing approach.

4.2 Hierarchical Trajectory Clustering Based PPM

4.2.1 Hierarchical PPM Framework

Spatial data can be expressed at different levels of detail (Haining, 2003). The notion of hierarchical spatial clusters supports hierarchical partitions of the space, by applying a hierarchical design to trajectory data. How to design a hierarchical method to explore and discover different levels of reference spots is a domain-specific task and worth exploring. Figure 4.3 shows an architecture of our proposed hierarchical PPM framework. First, we use spatio-temporal trajectory datasets with regular time intervals as input. Using regular time intervals is for period detection, and then single-level clustering and multi-level hierarchical clustering are applied for finding reference spots. In the process of single-level clustering, the framework implements two point based clustering approaches used in traditional spatio-temporal PPM: Kernel function (Li et al., 2012) and Grid-based (Giannotti et al., 2007), one trajectory based clustering Traclus (Lee et al., 2007) and our proposed algorithm. Note that, the point based clustering approaches ignore the sequence of trajectory nodes whilst the trajectory based clustering considers the sequence.

The single-linkage merge approach is the most popular merging technique in hierarchical clustering (Lee and Yang, 2009). In this method,

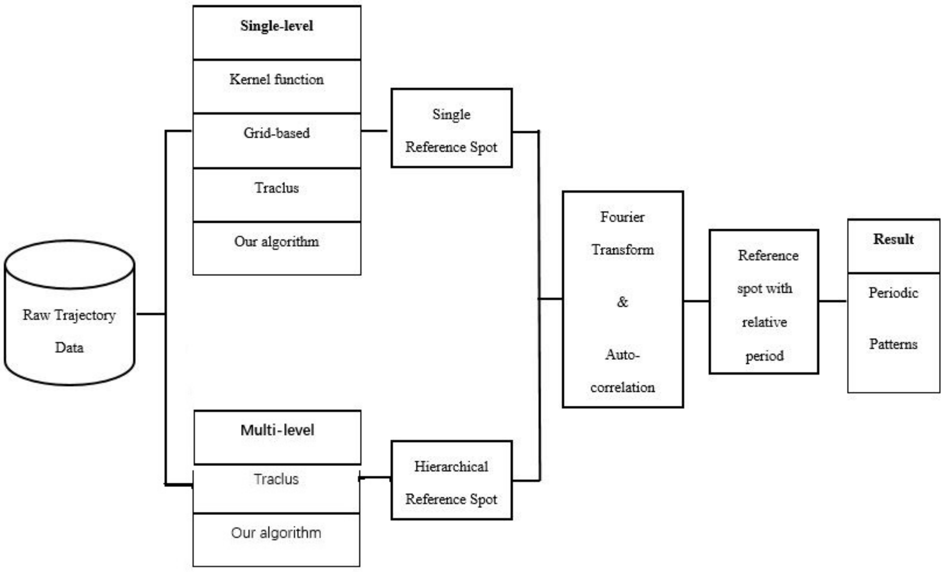


FIGURE 4.3: Framework of hierarchical PPM.

the distance between two clusters is the shortest distance between all pairs of patterns drawn from the two clusters (H.A. Sneath and R. Sokal, 1963). This chapter utilises the single-linkage with Traclus and our algorithm to find hierarchical reference spots. After gaining reference spots, we will apply Fourier transform and auto-correlation (Bar-David, 2009; Kargupta, 2005) to obtain periods for each reference spot. Finally, we apply the algorithm in (Li et al., 2012) to mine periodic patterns. Please refer to (Bar-David, 2009; Cao et al., 2007; Lee et al., 2007; Li et al., 2012; H.A. Sneath and R. Sokal, 1963; Kargupta, 2005) for details of approaches used in this chapter.

4.2.2 Hierarchical Trajectory Clustering Methodology

Our algorithm is based on Traclus (Lee et al., 2007). It is a trajectory clustering algorithm that considers the sequence of trajectory. This algorithm includes two phases for clustering trajectories: partition and group. It first partitions a trajectory into a set of line segments by characteristic points at which the object makes a sharp turn, and then, groups similar line segments together into a cluster. In the group phase, it extends traditional density-based points clustering algorithm DBSCAN to line-segment based clustering algorithm, which means it only considers the spatial attribute to calculate distance for clustering. Note that, traditional density-based clustering algorithms such as DBSCAN and DENCLUE (Hinneburg and Keim, 1998) cannot find hierarchical clusters and they fail to detect clusters of different densities since they use a global density threshold. In order to achieve hierarchical trajectory clustering for

PPM, we utilise HDBSCAN (Campello et al., 2013). It is an incremental version of DBSCAN which is able to handle clusters with different densities while preserving the scalability of DBSCAN. HDBSCAN is capable of generating a hierarchical clustering result using the single-linkage method, and then automatically extracts clusters from a hierarchical tree. We extend HDBSCAN to handle line segments and plug it into Traclus to detect hierarchical clusters.

Note that, trajecotry is a series of spatial locations with timestamps, thus it is explicitly spatio-temporal. Speed and direction properties are two implicit trajectory properties and they have been widely used in many trajectory data mining (Bermingham and Lee, 2015; Yuan et al., 2012; Zheng and Zhou, 2011). This chapter further extends Traclus with HDBSCAN to incorporate these two additional implicit trajectory properties. Thus, this quadruple $< Space, Direction, Speed, Time >$ covers both the explicit and implicit features of trajectory and captures more meaningful behaviors of trajectory.

As known, distance is a numerical description of how far apart objects are. In this chapter, we consider spatial distance, temporal distance, and additional directional distance and speed distance in order to consider spatio-temporal and sequential information.

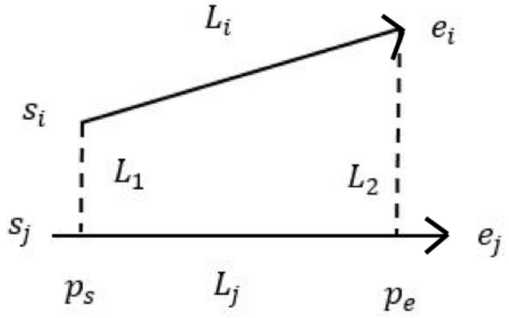


FIGURE 4.4: Spatial distance for two line segments $L_i = (s_i, e_i)$ and $L_j = (s_j, e_j)$ (p_s and p_e are the projection points from s_i and e_i onto L_j, respectively. $d(s_i, p_s)$ is the Euclidean distance between s_i and p_s, and $d(e_i, p_e)$ is that between e_i and p_e.).

Spatial distance ($SpaDist$) is to measure the geographic distance difference between trajectory segments. For example, two trajectory segments $L_i = (s_i, e_i)$ and $L_j = (s_j, e_j)$ as shown in Figure 4.4, the spatial distance is computed as below:

$$SpaDist(L_i, L_j) = \frac{d(s_i, p_s)^2 + d(e_i, p_e)^2}{d(s_i, p_s) + d(e_i, p_e)}, \tag{4.1}$$

where $d(s_i, p_s)$ is the Euclidean distance between s_i and p_s, whilst $d(e_i, p_e)$ is that between e_i and p_e.

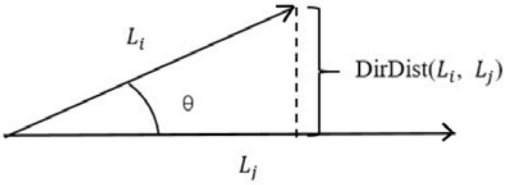


FIGURE 4.5: Directional distance for two line segments L_i and L_j (θ is the included angle between L_i and L_j).

Directional distance ($DirDist$) is to measure the difference in directional movements between trajectory segments. Given two trajectory segments L_i and L_j as shown in Figure 4.5, the directional distance is computed as below:

$$DirDist(L_i, L_j) = \begin{cases} min(||L_i||, ||L_j||) \times sin(\theta) & (0 \leq \theta \leq 90), \\ min(||L_i||, ||L_j||) \times sin(\pi - \theta) & (90 < \theta \leq 180), \end{cases} \tag{4.2}$$

where $||L_i||$ and $||L_j||$ are the length of L_i and L_j, respectively, and θ is the included angle between L_i and L_j. $DirDist(L_i, L_j)$ represents the deviation of direction in moving tendency between L_i and L_j (Yuan et al., 2012).

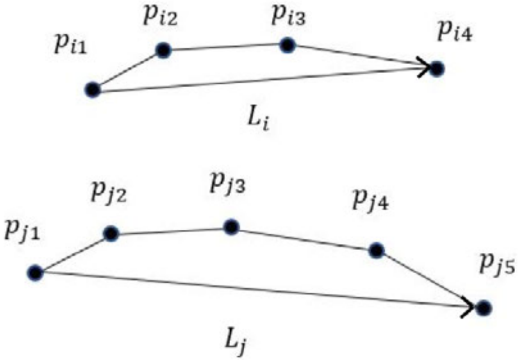


FIGURE 4.6: Speed distance for two line segments $L_i = (p_{i1}, p_{i4})$ and $L_j = (p_{j1}, p_{j5})$.

Speed distance ($SpdDist$) is to measure the difference in speed between trajectory segments. For two given line segments $L_i = (p_{i1}, p_{i4})$ and $L_j = (p_{j1}, p_{j5})$ as shown in Figure 4.6, $SpdDist(L_i, L_j)$ is calculated as below:

$$SpdDist(L_i, L_j) = \frac{S_{max}(L_i, L_j) + S_{avg}(L_i, L_j) + S_{min}(L_i, L_j)}{3}, \tag{4.3}$$

where $S_{max}(L_i, L_j)$ represents the absolute difference in maximum speed of two segments. Correspondingly, S_{avg} and S_{min} represent absolute differences in average and minimum speed of two segments. Note that, in the partitioning phase of Traclus, the whole trajectory is first examined to identify a sequence of characteristic points at which the object makes a sharp turn, and then the whole trajectory is partitioned into line segments by these characteristic points. In Figure 4.6, p_{i1} and p_{i4}, p_{j1} and p_{j5} can be represented as characteristic points, thus, L_i and L_j can be obtained.

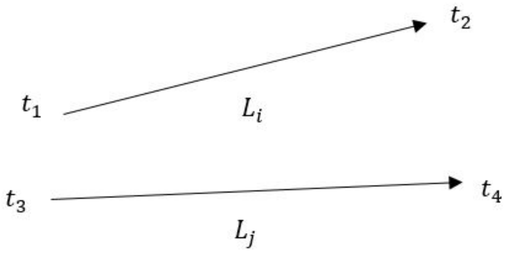


FIGURE 4.7: Two line segments L_i with two end points: (p_{i1}, t_{i1}) and (p_{i2}, t_{i2}), and L_j with two end points: (p_{j1}, t_{j1}) and (p_{j2}, t_{j2}), where p_k is a spatial location and t_k is a corresponding timestamp for $k \in \{i1, i2, j1, j2\}$.

For two given line segments $L_i = ((p_{i1}, t_{i1}), (p_{i2}, t_{i2}))$ and $L_j = ((p_{j1}, t_{j1}), (p_{j2}, t_{j2}))$ where p_k is a spatial location and t_k is a corresponding timestamp for $k \in \{i1, i2, j1, j2\}$ as shown in Figure 4.7, time distance ($TimeDist$) is to measure the difference during time intervals between trajectory segments. $TimeDist(L_i, L_j)$ is represented by:

$$TimeDist(L_i, L_j) = \sqrt{(t_{j1} - t_{i1})^2 + (t_{j2} - t_{i2})^2}. \tag{4.4}$$

These four distance measures have different scales. In order to avoid the scaling problem, we use the min-max normalisation (Han, 2005) to scale each distance measure into the unit range [0,1]. The total distance ($TotalDist$) is now defined as below:

$$\begin{aligned} &TotalDist(L_i, L_j) = \\ &\qquad W_{spa} * SpaDist\prime(L_i, L_j) + W_{dir} * DirDist\prime(L_i, L_j) \\ &\qquad\quad + W_{spd} * SpdDist\prime(L_i, L_j) + W_{time} * TimeDist\prime(L_i, L_j), \end{aligned} \tag{4.5}$$

where $SpaDist\prime$, $DirDist\prime$, $SpdDist\prime$ and $TimeDist\prime$ are min-max normalised distances whilst W_{spa}, W_{dir}, W_{spd} and W_{time} are relative weights for $SpaDist\prime$, $DirDist\prime$, $SpdDist\prime$ and $TimeDist\prime$, respectively. Note that, the sum of weights is equal to 1 (W_{spa}+W_{dir}+W_{spd}+W_{time}=1). In this chapter, we use the same equal weight for these distances but they can vary with applications. That is, W_{spa}=W_{dir}=W_{spd}=W_{time}=0.25 in this experiment. This $TotalDist(L_i, L_j)$ replaces the distance function used in Traclus.

Note that, the quadruple $< Space, Direction, Speed, Time >$ can represent the implicit and explicit properties of trajectory and capture more meaningful trajectory behaviours. Namely, each property expresses periodicity: people go to office on weekdays (periodic spatial pattern); people keep similar speed in different road segments (periodic speed pattern); people move north-west-east-north of city (periodic directional pattern).

4.3 Experimental Results

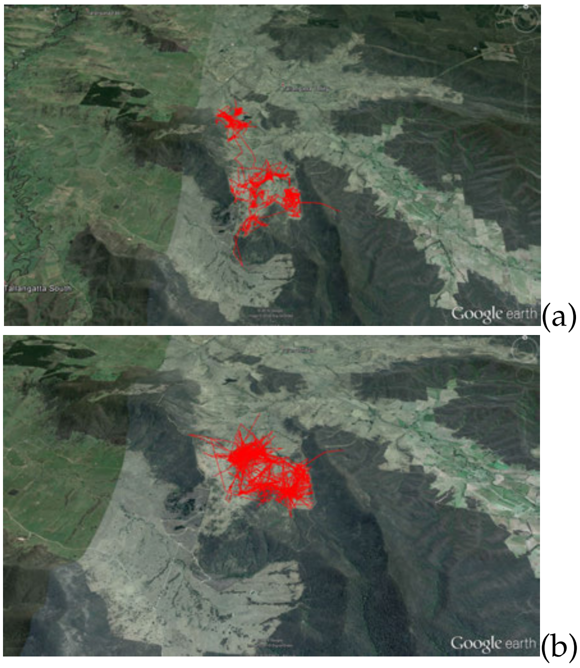


FIGURE 4.8: Real world datasets: (a) Dataset 1: Free-ranging Maremma sheepdogs dataset; (b) Dataset 2: GPS collar dataset.

TABLE 4.1: Format of Dataset 1 and 2.

tag-local-identifier	Longitude	Latitude	Timestamp	Sensor-type
1	147.367458	36.318158	2012-07-30,00:35:18	gps

Our aim is to find quality reference spots with non-zero periods as possible so that more periodic patterns can be found in order not to miss false positives. Experiments are based on two real datasets from movebank. [1] Movebank is a free online animal tracking database which helps researchers to share and analyse animal movement data. The database is designed to capture individual animal tracking data. Information from animal movements is significant for movement ecology, such as climate change, disease spread and biodiversity loss. The main aims of Movebank include: (1) to archive animal movement data; (2) to help scientists explore new questions by combining these datasets; (3) to allow the public to explore these datasets. We use GPS tracking collar data which describe

[1] https://www.movebank.org

movements and behaviours of free-ranging livestock guardian dogs on three properties in Victoria, Australia. We use Google Earth[1] to visualise the datasets. The first dataset (Figure 4.8(a)) contains around 4 months long tracking records (7/2012 - 11/2012), while the second one (Figure 4.8(b)) describes more than 2 months long trajectories (4/2011 - 6/2011). Table 4.1 shows a part of attributes in two datasets. The tag-local-identifier presents id which can identify different moving objects. Location is shown with longitude and latitude. Timestamp means recorded time in the location. Sensor-type represents that the datasets are GPS recorded. In our study, we choose two single moving objects and their location and corresponding timestamp. Note that, we need some interpolation methods to make these data regular time intervals. In this chapter, we use the cubic interpolation (McKinley and Levine, 1998). Since most of the nodes have the same time interval (30 minutes), there is minimal influence from this interpolation preprocessing.

4.3.1 Dataset 1

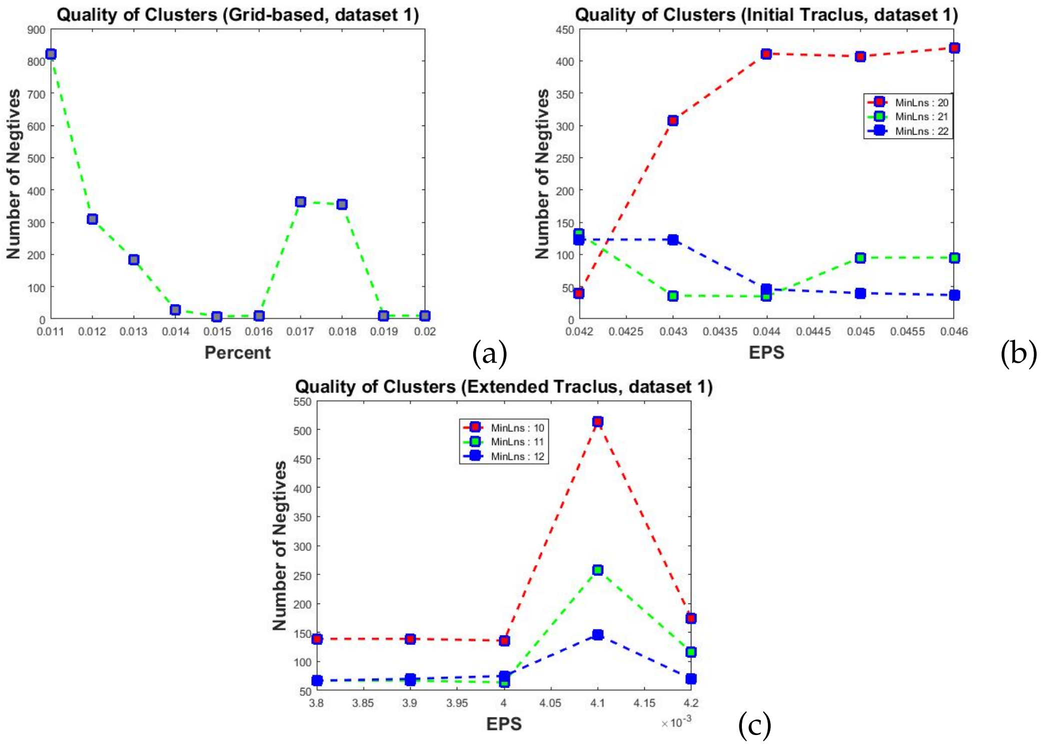


FIGURE 4.9: Selection of parameter values for Dataset 1: (a) Grid-based; (b) Initial Traclus; (c) Extended Traclus.

[1] https://www.google.com/earth

Selection of Parameter Values

As most data mining approaches require parameter-tuning (Han, 2005), the clustering approaches under study in this chapter also require parameter-tuning to produce best possible results. The main aim of this subsection is to empirically evaluate parameter values for the clustering approaches used in this chapter. Figure 4.9 shows the measure of cluster quality, and we use the silhouette coefficient (Rousseeuw, 1987) for this aim. The silhouette value ranges from -1 to +1. A high silhouette value indicates that a cluster is well-matched to its own cluster, and poorly-matched to neighboring clusters. If most points have a high silhouette value, then the clustering solution is appropriate. If many points have a low or negative silhouette value, then the clustering solution may have either too many or too few clusters. Figure 4.9 displays the number of negatives in Y-axis while presenting parameter values for corresponding clustering algorithms in X-axis. Thus, a lower value in Y-axis results in a high silhouette value. Figure 4.9(a) shows the quality measure for Grid-based clustering. We set the density threshold from 1% to 2% of total number of data points in X-axis, and Y-axis presents the number of negatives. There is a peak when the density threshold is 1.1%, which means the clustering solution is inappropriate. Oppositely, after 1.5%, the clustering becomes stable and better, thus, we choose 1.5% as a dense threshold for the first dataset. Figure 4.9(b) shows the quality measure for original initial Traclus. We can get an optimal value when ε = 0.044 and $MinLns$ = 21. Figure 4.9(c) displays the quality measure of our extended Traclus, and the optimal value can be set to ε = 0.004 and $MinLns$ = 11. For Kernel function, we use a top 15% density value threshold for cluster, most parts of reference spots should be detected with high probability (Li et al., 2012).

Comparison among Clustering Methods

In this part, we undertake comparative studies among three existing clustering algorithms and our algorithm for Dataset 1.

Figure 4.10 shows visualisations of reference spots using three clustering methods and our extended Traclus for Dataset 1. Numbers in red rectangles represent *i*-th reference spots. We use different colors to distinguish different reference spots for Traclus-based approaches which include initial Traclus and our method. Figure 4.10(a) shows the result of Kernel Function finding 4 reference spots (the areas with red plus). Kernel Function finds the least number of reference spots, and only returns globally high dense reference spots. This approach misses many other locally formulated dense reference spots as shown in Figure 4.10(a). Figure 4.10(b) displays 7 reference spots found with Grid-based whilst Figure 4.10(c) exhibits 7 reference spots (lines with different color) with inital Traclus. Figure 4.10(d) illustrates 7 reference spots detected with proposed extended Traclus. Initial Traclus and extended Traclus find more reference spots than Kernel Function in single-level. They return global and local dense reference spots that will be used for interesting periodic patterns.

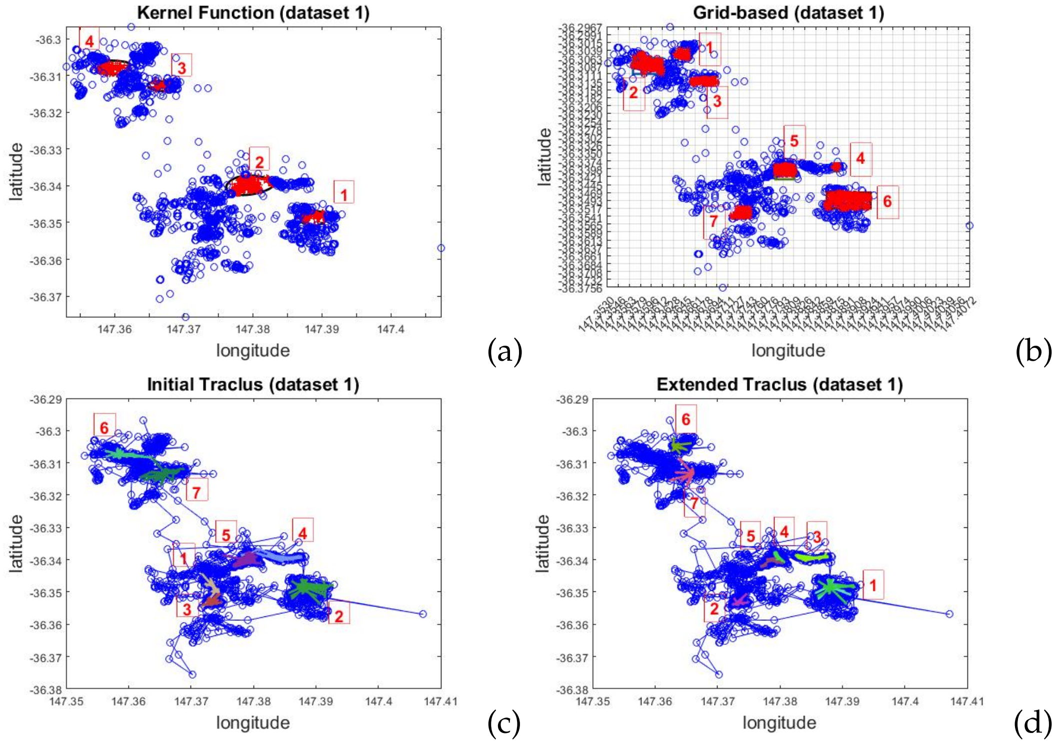


FIGURE 4.10: Four clustering results for Dataset 1: (a) Kernel Function (4 reference spots); (b) Grid-based (7 reference spots); (c) Traclus (7 reference spots); (d) Extended Traclus (space+time, 7 reference spots).

TABLE 4.2: Number of reference spots found from three existing clustering in single-level for Dataset 1. (RS: the number of Reference Spots; NZRS: the number of Reference Spots with Non-Zero period).

	Kernel Function	Grid-based	Initial Traclus
RS	4	7	7
NZRS	2	2	5

Table 4.2 represents the number of reference spots, and the number of reference spots with non-zero period. Note that, we only choose reference spots with non-zero period to obtain periodic patterns in the later phase in this chapter. As we have mentioned before, the aim of PPM is to find regularly repeating activities for reference spots. Oppositely, if a reference spot with 0 period, it would represent that an object goes to this place with irregular time or infrequent. Thus, RS is useful but NZRS is more useful and meaningful to mine periodic patterns. Also, it is worth noting that Grid-based has the same reference spots with initial Traclus, but the latter obtains more reference spots with non-zero period, which means more periodic patterns can be found from initial Traclus compared to Grid-based.

TABLE 4.3: Reference spots from extended Traclus in single-level for Dataset 1 (Sc: Space; Di: Direction; Sd: Speed; Ti: Time).

	ScDi	ScSd	ScTi	ScDi Sd	ScDi Ti	ScSd Ti	ScDi SdTi
RS	6	18	7	8	6	7	7
NZRS	4	9	3	4	2	3	3

Table 4.3 displays the number of reference spots for various combinations of the four attributes considered (space, direction, speed and time) for our proposed method. Obviously, we can find useful and hidden periodic patterns in all combinations. For instance, in Figure 4.11, we obtain 8 reference spots considering the combination of space, direction and speed, and then a period (147 hours) at 6th reference spot is found. However, three existing methods cannot detect this reference spot and thus cannot obtain a non-zero period for that. Comparing Table 4.2 and Table 4.3, extended Traclus can find more RS and NZRS, which means more interesting periodic patterns can be found.

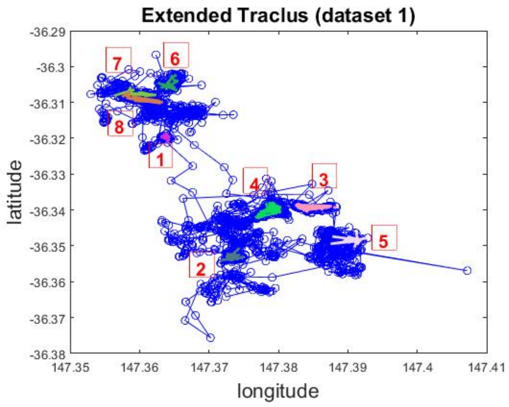


FIGURE 4.11: Extended Traclus (Space+Direction+Speed) results for Dataset 1.

Hierarchical Clustering

In this section, initial Traclus and extended Traclus are extended to consider hierarchy using HDBSCAN with the single-linkage approach in order to find hierarchical reference spots.

Figure 4.12 represents corresponding dendrograms obtained using HDBSCAN with the single-linkage algorithm. Dendrograms can be broken at different levels to yield different clusterings of data (Jain and Dubes, 1988). For instance, Figure 4.12(a) represents a dendrogram of Dataset 1 for initial Traclus. Clusters 1 and 3 can be combined as a new reference spot, and also another new reference spot can be formed by merging clusters 4 and 5. Figure 4.12(b) shows a dendrogram of Dataset 1 for extended Traclus (considering space and time).

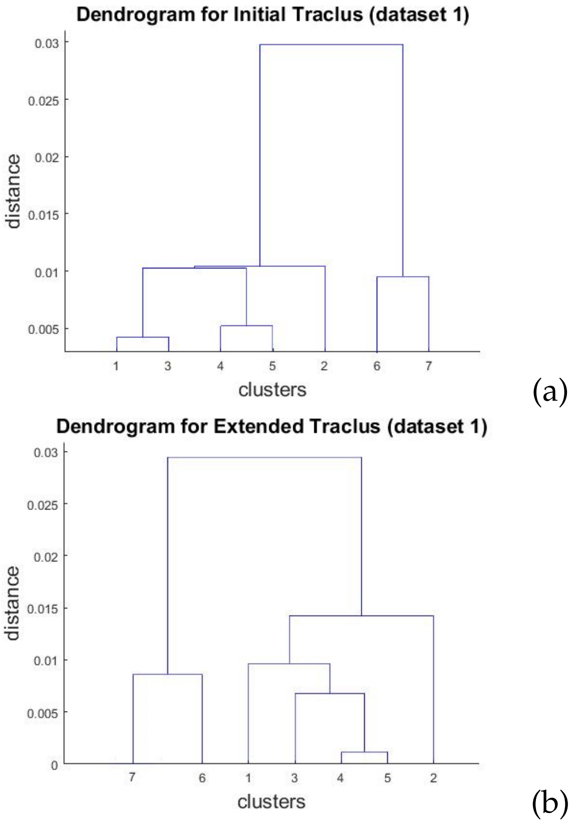


FIGURE 4.12: Dendrograms for initial Traclus and extended Traclus clustering algorithms for Dataset 1: (a) Initial Traclus; (b) Extended Traclus (space+time).

TABLE 4.4: Hierarchical reference spots with initial Traclus for Dataset 1.

	Initial Traclus
RS	6
NZRS	5

Table 4.4 displays the number of hierarchical reference spots with initial Traclus. Obviously, more periodic patterns can be mined with this hierarchical initial Traclus. For instance, Figure 4.13 shows reference spots 1 and 3 (shown in Figure 4.10(c)) can be combined to reference spot 8, and then, we can mine some periodic patterns for reference spot 8. This hierarchical reference spot 8 could represent an hierarchical geographical place type that could not be detected by the single-level clustering approaches.

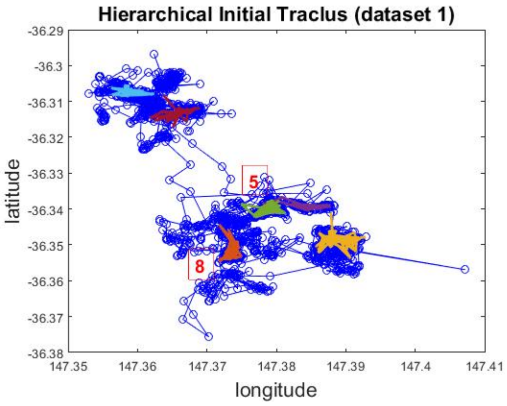


FIGURE 4.13: Hierarchical initial Traclus for Dataset 1.

TABLE 4.5: Reference spots for hierarchical extended Traclus for Dataset 1.

	ScDi	ScSd	ScTi	ScDi Sd	ScDi Ti	ScSd Ti	ScDi SdTi
RS	5	17	6	7	5	6	6
NZRS	4	14	5	6	4	4	5

As shown in Table 4.5, various reference spots are found with various combinations in the hierarchy. For instance, Figure 4.14 displays reference spots 4 and 5 (shown in Figure 4.10(d)) are combined as a larger one. More periodic patterns can be mined for reference spot 8.

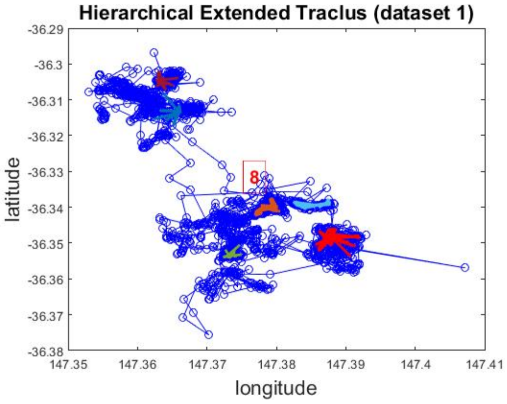


FIGURE 4.14: Hierarchical extended Traclus (space+time) for Dataset 1.

In our proposed framework, we allow users to take additional information into account in clustering including direction, speed and time in addition to the basic spatio-temporal information. Choosing additional semantic information is context-dependent (Ilarri, 2011), and we provide experimental results with all different combinations of these additional semantic information to provide users with various reference spots for their decision-making. Note that, since we are dealing with spatio-temporal trajectories, 'ScTi' considering space and time is used as default in this chapter.

Table 4.3 and Table 4.5 display the number of reference spots for various combinations of the four attributes considered (space, direction, speed and time). In all combinations, the number of NZRS generated with hierarchical clustering exceeds that of NZRS generated with corresponding single-level clustering.

4.3.2 Dataset 2

Selection of Parameter Values

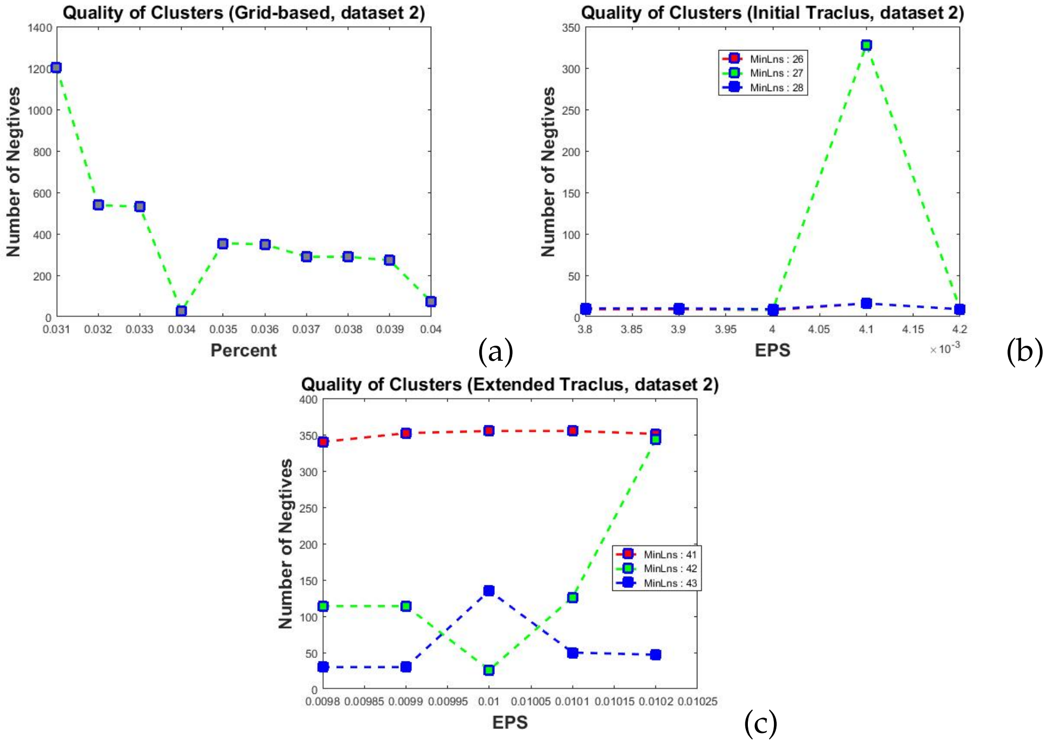


FIGURE 4.15: Selection of parameter values for Dataset 2: (a) Grid-based; (b) InitalTraclus; (c) Extended Traclus (space+time).

Similar to Section 4.1.1, Figure 4.15 shows the quality measure for Dataset 2. Figure 4.15(a) shows the quality measure for Grid-based where we choose 3.4% as the density threshold. The quality measure for Traclus is shown in Figure 4.15(b). $\varepsilon = 0.004$ ($MinLns = 27$) has the lowest number of negatives (a high silhouette value), and it obtains more reference spots. Thus, ($\varepsilon = 0.004$, $MinLns = 27$) are better values gaining a higher silhouette value and more reference spots. In Figure 4.15(d), obviously, ($\varepsilon = 0.01$, $MinLns = 42$) have the lowest value. For Kernel function, we still use the top-15% density value threshold for clustering.

Comparison among Three Existing Clustering Methods

For Dataset 2, Figure 4.16 shows visualisations of reference spots for each clustering method. Figure 4.16(a) shows the clustering result of Kernel Function with 2 reference spots, while Figure 4.16(b), (c), (d), Grid-based, inital Traclus and extended Traclus depict the same number of reference spots.

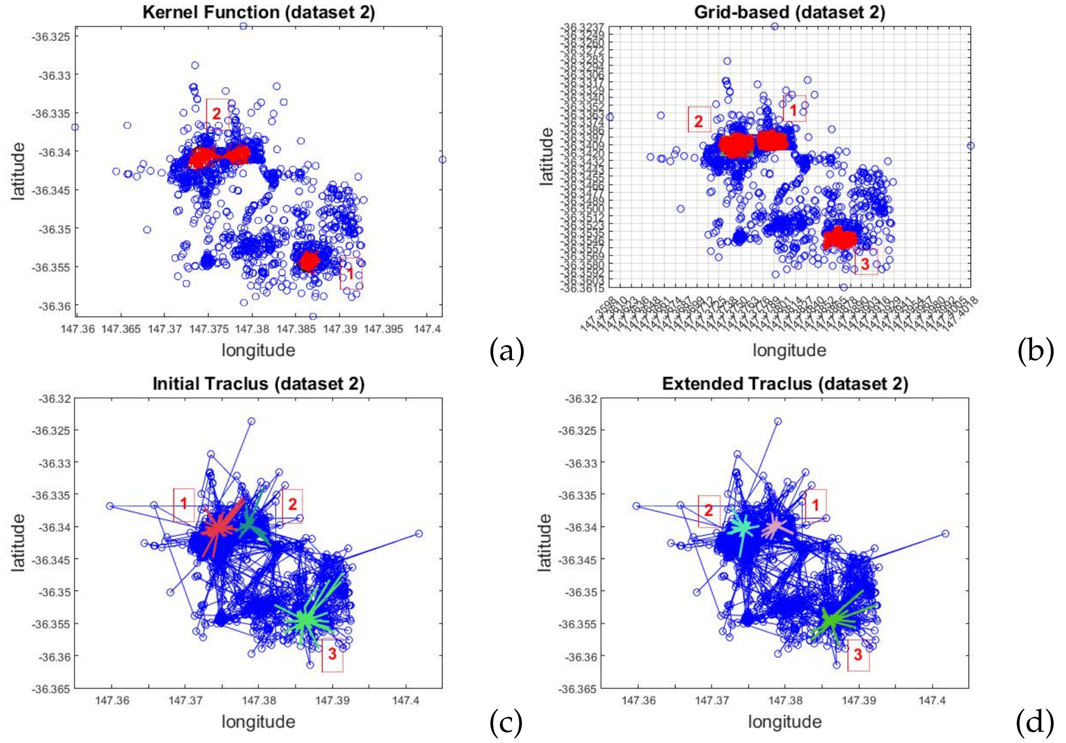


FIGURE 4.16: Four clustering results for Dataset 2: (a) Kernel Function (2 reference spots); (b) Grid-based (3 reference spots); (c) Traclus (3 reference spots); (d) Extended Traclus (space+time, 3 reference spots).

TABLE 4.6: Reference spots from three existing clustering methods in single-level for Dataset 2.

	Kernel Function	Grid-based	Initial Traclus
RS	2	3	3
NZRS	2	2	3

Similar to Dataset 1, in Table 4.6 and 4.7, the sequence based clustering methods (initial Traclus and extended Traclus) produce more NZRS than corresponding point based approaches (Kernel Function and Grid-based) with Dataset 2, and different combinations of these additional semantic information can produce more reference spots. In addition, hierarchical clustering produces more NZRS than corresponding single-level clustering with Dataset 2. Table 4.8 and Table 4.9 confirm these with Dataset 2.

TABLE 4.7: Reference spots from extended Traclus (our algorithm) for Dataset 2.

	ScDi	ScSd	ScTi	ScDi	ScDi	ScSd	ScDi
				Sd	Ti	Ti	SdTi
RS	3	6	3	4	3	3	3
NZRS	1	3	1	2	1	1	3

Hierarchical Clustering

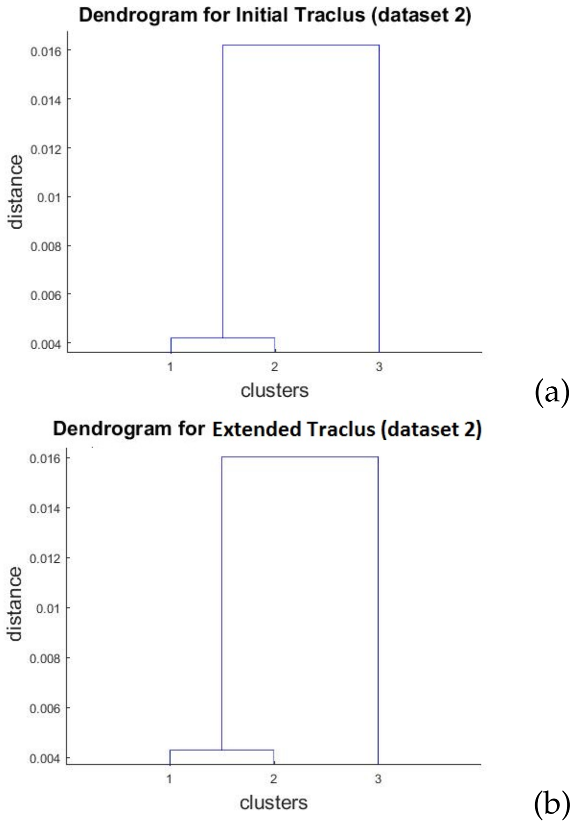


FIGURE 4.17: Dendrograms for initial Traclus and extended Traclus for Dataset 2: (a) Initial Traclus; (b) Extended Traclus (space+time).

Figure 4.17 shows dendrograms of hierarchical clustering with Dataset 2. Similar to Dataset 1, initial Traclus and extended Traclus reveal many interesting hierarchical reference spots.

TABLE 4.8: Reference spots for hierarchical initial Traclus for Dataset 2.

	Initial Traclus
RS	3
NZRS	2

TABLE 4.9: Reference spots for hierarchical clustering for Dataset 2 (our algorithm).

	ScDi	ScSd	ScTi	ScDi Sd	ScDi Ti	ScSd Ti	ScDi SdTi
RS	3	5	3	3	3	3	3
NZRS	2	4	2	2	2	2	2

4.3.3 Effectiveness and Efficiency

Effectiveness

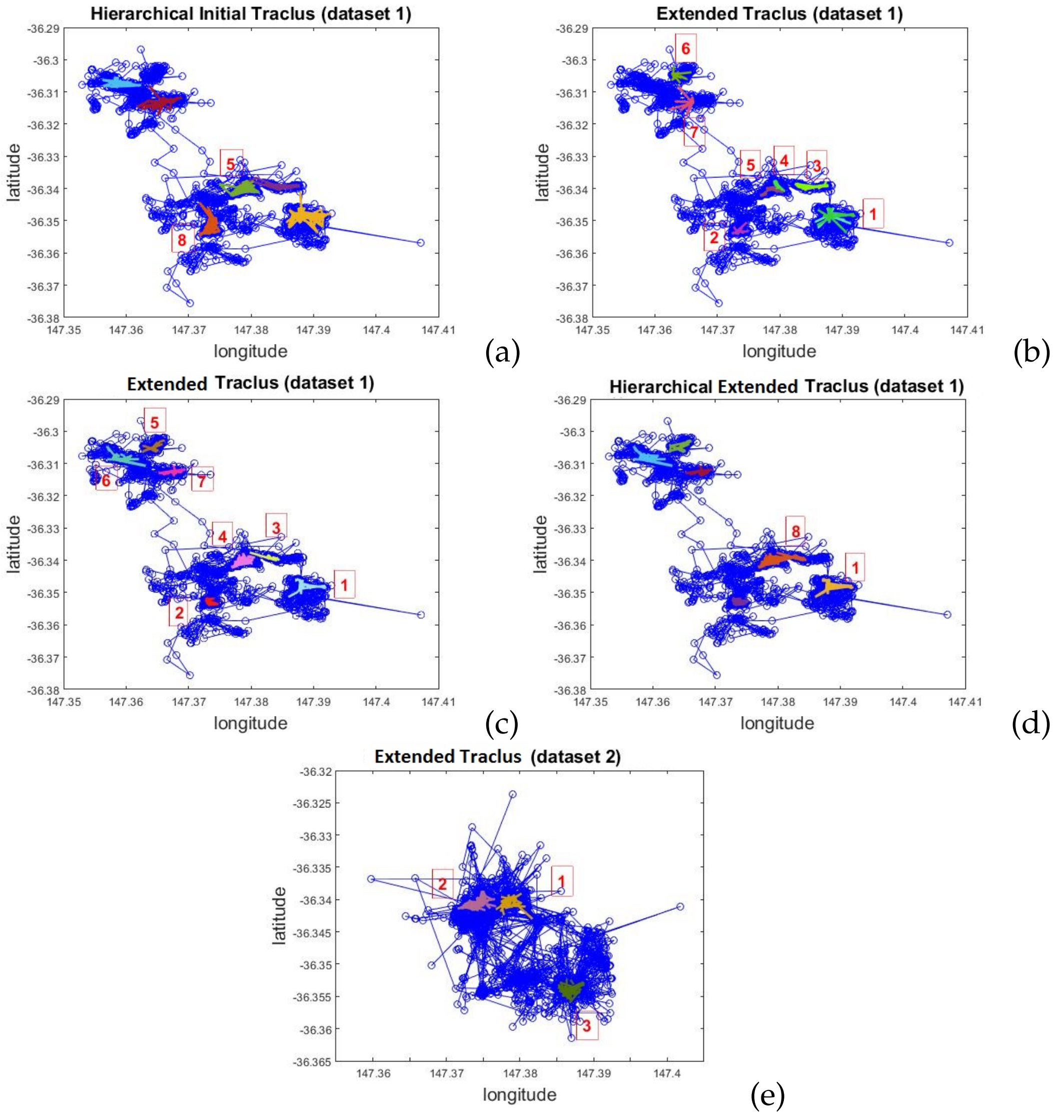


FIGURE 4.18: Some interesting periodic patterns for Datasets 1 and 2: (a) Hierarchical initial Traclus; (b) Single-level extended Traclus (space+time); (c) Single-level extended Traclus (all 4 properties); (d) Hierarchical extended Traclus (all 4 properties); e) Single-level extended Traclus (all 4 properties).

We have listed RS and NZRS, but we will use NZRS for PPM method to extract periodic behaviors and patterns. Due to the page limit, we list some periodic patterns. Figure 4.18 visualises interesting reference spots detected by our approach.

TABLE 4.10: Number of periodic behaviors (patterns).

	Reference spot	period (Hours)	periodic behaviors	Periodic Patterns
Figure 4.18(a)	8	67.5	7	5→0→8
Figure 4.18(b)	3	47.5	2	0→3→0
Figure 4.18(c)	6	239.5	2	5→0→6
Figure 4.18(c)	6	166.5	3	6→0→7
Figure 4.18(d)	8	22	16	1→0→8
Figure 4.18(e)	1	24.5	4	2→0→1
Figure 4.18(e)	2	25.5	20	1→0→2

Table 4.10 shows some periodic patterns from reference spots shown in Figure 4.18. Note that, traditional approaches are not able to detect these periodic behaviors. Figure 4.18(a) displays hierarchical reference spots identified by initial Traclus, it shows that reference spot 8 is periodically visited with a regular time interval (67.5 hours), from reference spot 5. In reality, this happens frequently, for instance, Bob picks up his best friend before going to work sometimes, the place where his friend is living should be related to their work place. Thus, in Figure 4.18(a), we can say that there is a periodic pattern from reference spot 5 to 8 with 67.5 hours. There are 7 periodic behaviors for Figure 4.18(a), which means a person arrives at reference spot 8 at different time. For instance, Bob arrives at office at 8:00 on Moday, at 9:00 on Tuesday, and at 8:00 on Wednesday etc. Figure 4.18(b) shows those identified by the space and time combination. In the same period, we can find a periodic pattern between somewhere (0 means not in any reference spot) and reference spot 3 with 47.5 hours, and 2 periodic behaviors. Figure 4.18(c) displays two periodic patterns, 5→0→6 and 6→0→7, it does not necessarily mean that we can get 5→0→6→0→7 since two patterns do not frequently occur together. Figure 4.18(d) reveals a periodic pattern 1→0→8 (22 hours) between hierarchical reference spots 1 and 8. In Figure 4.18(e), we can find that reference spots 1 and 2 have a similar period, which is around 25 hours, periodic pattern, 2→0→1 is for reference spot 1, while 1→0→2 is for reference spot 2, which means the object moves periodically between reference spots 1 and 2, and 20 different periodic behaviors from reference spot 1 to reference spot 2 in a period. In reality, reference spot 1 and 2 can be seen as work place and home.

Efficiency

Figure 4.19 shows the running time analysis of the proposed trajectory clustering between single-level and multi-level with Dataset 1. When compared to single-level, hierarchy does not require a great amount of extra time, but affordable additional time.

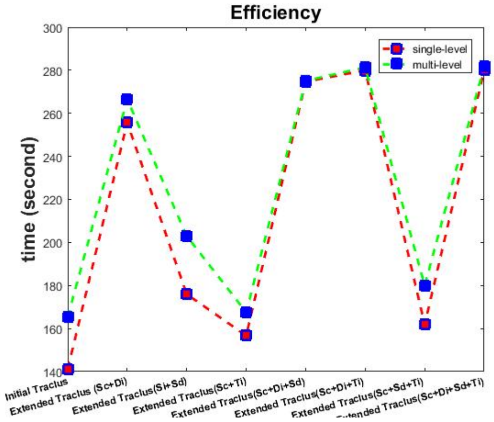


FIGURE 4.19: Efficiency for single-level and multi-level for Dataset 1.

4.4 Summary

In this chapter, we identify four crucial drawbacks of traditional PPM methods: disregarding the sequence of trajectory, ignoring hierarchical nature of clustering, not consider spatio-temporal aspects at the same time, and not considering trajectory path. We propose a new path-based clustering method based on Traclus to find reference spots (trajectory paths). We utilise a trajectory clustering approach in order to consider the sequential nature of trajectory and also use a hierarchical clustering method to generate hierarchical reference spots. Experimental results show that our hierarchical approach requires slightly more time than the single-level approach with a small margin, but generates more reference spots and periodic patterns that traditional approaches are unable to detect.

Chapter 5

Hierarchical Semantic PPM from Spatio-temoral Trajectories

In this chapter, we present the node-based approach with consideration of semantic background information for PPM from spatio-temporal trajectory. Section 5.1 represents an introduction of this study. Section 5.2 shows the overall framework and presents the details of node-based approach. The experimental results are represented and discussed in Section 5.3. In Section 5.4, we present and discuss the experimental results for hierarchical semantic reference spots and hierarchical periodic patterns. In addition, we compare this to the path-based approach to demonstrate that our method in this chapter (node-based approach) outperforms Traclus (ST) (path-based approach) when we consider semantic background information. Finally, Section 5.5 shows conclusive remarks.

5.1 Introduction

Periodic patterns exhibit important repeating and regular behaviours of a moving object at a certain place. According to the literature, existing PPM approaches for spatio-temporal trajectories suffer from different problems. Note that these GPS-collected trajectory datasets represent real-world movement phenomena and thus they are sequentially connected, spatially placed, temporally recorded, aspatial semantically meaningful, hierarchically structured, and irregularly sampled. Therefore, PPM from spatio-temporal trajectories must consider the following special characteristics of GPS collected spatio-temporal trajectories. They are: (1) consideration of trajectory sequence in reference spot detection. Spatio-temporal trajectory is a sequence which is comprised of successively sampled points. Each sampled point is associated with the previous sampled point and the next one; (2) simultaneous consideration of spatiality and temporality features in reference spot detection. A trajectory is a temporal sequence of spatial locations where each location is associated with a timestamp describing the movement of an object. Thus, more effective periodic patterns can be mined when space and time are considered simultaneously; (3) consideration of background semantic information in reference spot detection. Note that spatio-temporal trajectories capture movements of real-world objects, and real-world phenomena have three dimensions: spatiality, temporality and aspatial semantics (Ying et al., 2011). Thus considering the aspatial semantic dimension in spatio-temporal PPM is of importance in order not to miss any spatially, temporally and semantically meaningful periodic patterns; (4) consideration of hierarchical nature of spatio-temporal data due to the inherent hierarchy of space. More hidden periodic patterns can be mined when the hierarchical nature of space is considered; (5) consideration of handling irregular spatio-temporal trajectory data for period detection. In reality, it is highly impossible that spatio-temporal trajectory data can be obtained with regular time intervals due to weather conditions, battery issues, and device malfunctions. Therefore, robust PPM from spatio-temporal trajectories must be able to manage this irregularity of spatio-temporal trajectories.

In Chapter 4, we have proposed a path-based approach called Traclus (ST), which considers both spatial and temporal aspects at the same time in order to find spatio-temporal concentrations when mining (hierarchical) periodic patterns. Although this approach can effectively obtain periodic patterns based on trajectory paths considering the sequence of trajectory, the hierarchy of space and two important dimensions (spatial and temporal), it disregards another important dimension aspatial semantic dimension (descriptive geographical feature information). Spatio-temporal trajectories are geographical phenomena occurring in the geographical space. Geographical phenomena have three dimensions: spatial, temporal and aspatial semantic (Peuquet, 2002; Ying et al., 2011), thus considering the aspatial semantic dimension in spatio-temporal trajectory mining is of importance in order not to miss any spatially, temporally and semantically meaningful periodic patterns.

All these traditional approaches assume that GPS trajectories are regular trajectories that are characterised by a constant time lag between two successive recordings. That is, they assume spatio-temporal trajectories are regularly sampled, and they take a regularly sampled trajectory data as input. In the real world, spatio-temporal trajectories are irregularly sampled due to weather conditions, device malfunctions, battery issues, bandwidth limitations and power issues (Li et al., 2016). These traditional approaches require a computationally expensive trajectory interpolation method (Bermingham and Lee, 2017) to make trajectory data regular to employ Fourier Transform and Autocorrelation (FT&Auto) in order to find regular periods for references spots.

In this chapter, we propose a node-based hierarchical semantic PPM from spatio-temporal trajectories. We first detect sequentially, spatially, temporally, semantically and hierarchically aggregated concentrations from irregular trajectories using aspatial semantic information from OpenStreetMap, and compute regular periodic time intervals for these spatio-temporal concentrations from irregular trajectories.

5.2 Framework and Algorithm

5.2.1 Framework and Algorithm

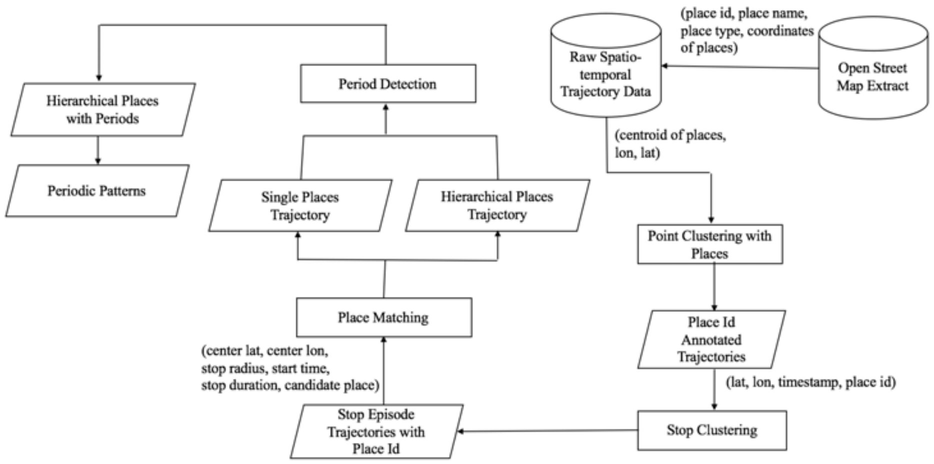

Hierarchical Places with Periods
Period Detection
Raw Spatio-temporal Trajectory Data
(place id, place name, place type, coordinates of places)
Open Street Map Extract
Periodic Patterns
Single Places Trajectory
Hierarchical Places Trajectory
(centroid of places, lon, lat)
Point Clustering with Places
Place Matching
Place Id Annotated Trajectories
(center lat, center lon, stop radius, start time, stop duration, candidate place)
(lat, lon, timestamp, place id)
Stop Episode Trajectories with Place Id
Stop Clustering

FIGURE 5.1: Framework of semantic PPM from spatio-temporal trajectories.

Figure 5.1 illustrates an architecture of our proposed framework for hierarchical semantic PPM from spatio-temporal trajectories. First, we extract geographical background semantic information from OpenStreetMap. This background information includes *place id*, *place name*, *place type* and *geometry*. Second, we compute centroids of places using corresponding geometries of places, and then add them to enrich raw spatio-temporal trajectories. Third, traditional clustering algorithm DBSCAN is employed using centroids of places as core points to find clusters for each place. The output is a semantic trajectory with annotated place id. In this step, we can also implement HDBSCAN to replace DBSCAN using the centroids of places as core points to generate hierarchical reference spots. The output of this process is a set of trajectories with annotated place IDs. Fourth, spatio-temporally aggregated trajectory nodes with annotated place id might indicate stops where a meaningful activity occurs. We cluster these possible stops to form stop episodes based on a user-specified time threshold. That is, if a user stays in a certain geographical place (stop) for more than the user-specified threshold, then we form a stop episode that indicates a meaningful activity in a certain geographical place. Fifth, each extracted stop episode is matched to a semantic place using Hidden Markov Model (HMM). Sixth, we apply Lomb-Scargle periodogram to detect regular periods from unevenly sampled data for each hierarchical semantic place. Finally, we mine semantic periodic patterns from single-level relevant places or multi-level (hierarchical) relevant places in a dendrogram, using the single-linkage approach merging two semantic places with the smallest distance.

Obviously, there are three key steps in our method, which include stop episode detection, place matching and period detection. After stop episode detection, we match each stop episode to relative semantic places, and then we can calculate the period for related semantic places. Algorithm 1 shows the overall procedure of our proposed semantic PPM from spatio-temporal trajectories for single-level. Algorithm 2 shows the pseudocode of our proposed semantic PPM from spatio-temporal trajectories for multi-level.

Algorithm 1 Semantic PPM from spatio-temporal trajectories

```
INPUT: A spatio-temporal trajectory $T$, and semantic background
    information (.OSM file);
OUTPUT: A set of periodic patterns;
/* Divide all points into cluster points */
Extracts centroids for places;
Apply DBSCAN to get clusters $C = \{c_1, c_2, \dots, c_k\}$ from $T$ and centroids
    for places;
/* Find stop episode */
for each $s_i \in c_i$ do
    $S = S \cup s_i$
end for
Find stop episodes $S = \{s_1, s_2, ..., s_n\}$;
/* Match stop episodes to places */
for each $s_i \in S$ do
    Match each stop episode in $S$ to places $P = \{p_1, p_2, ..., p_n\}$;
end for
/* Detect periods */
for each stop episode $s_i \in S$ do
    Detect periods for place $p_i$ that matches with $s_i$, and store the periods
    in $T_i$;
end for
/* Find periodic patterns */
for each $t \in T_i$ do
    $p_t = \{p_i \mid t \in T_i\}$;
    Construct a symbolised sequence $Q$ using $p_t$;
    Mining periodic patterns from $Q$;
end for
```

5.2.2 Stop Episode Detection

In this section, we give details of how to find stop episodes. First, we use OpenStreetMap to generate an OSM file. Relevant spatial and aspatial semantic information can be extracted from this file, including place ids, place names, place types and coordinates of places, and road networks. Second, we employ DBSCAN using the centroids of places as core points in raw spatio-temporal trajectories in order to find trajectory nodes (GPS points) which are close to places or within the places. The output is a semantic trajectory with annotated place id. One thing to note is that some trajectory nodes could have more than one place id assigned. Figure 5.2

Algorithm 2 Hierarchical semantic PPM from spatio-temporal trajectories

```
INPUT: A spatio-temporal trajectory $T$, and semantic background
    information (.OSM file);
OUTPUT: A set of semantic periodic patterns;
/* Divide all points into cluster points */
Extracts centroids for places;
Apply HDBSCAN to get clusters $HC = \{hc_1, hc_2, \ldots, hc_k\}$ from $T$ and
    centroids for places;
/* Find stop episode */
for each $s_i \in hc_i$ do
    $S = S \cup s_i$
end for
Find stop episodes $S = \{s_1, s_2, ..., s_n\}$;
/* Match stop episodes to places */
for each $s_i \in S$ do
    Match each stop episode in $S$ to places $P = \{p_1, p_2, ..., p_n\}$;
end for
/* Generate hierarchical reference spots */
Build hierarchical semantic places $HR = \{hr_1, hr_2, \ldots, hr_k\}$ from
    dendrogram;
/* Detect periods */
for each hierarchical semantic place $hr_i \in HR$ do
    Detect periods for each semantic place $hr_i$, and store the periods in $T_i$;
end for
/* Find periodic patterns */
for each $t \in T_i$ do
    $p_t = \{p_i \mid t \in T_i\}$;
    Construct a symbolised sequence $Q$ using $p_t$;
    Mining periodic patterns from $Q$;
end for
```

illustrates an example. It displays that there is a trajectory $T = \{p_1, p_2, \ldots, p_{14}\}$ through three buildings A1, A2 and A3.

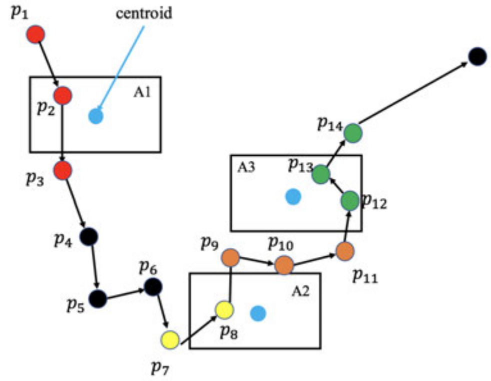


FIGURE 5.2: A spatio-temporal trajectory with stop or move annotations (nodes in black for moves otherwise stops).

After DBSCAN, some possible stops can be found, such as red points p_1, p_2 and p_3 belonging to A1, yellow and brown points p_7, p_8, p_9, p_{10} and p_{11} belonging to A2. Brown and green points p_9, p_{10}, p_{11}, p_{12}, p_{13} and p_{14} belonging to A3. Note that p_9, p_{10} and p_{11} are shared by A2 and A3. Thus, the raw trajectory can be represented as (*lon*$_1$, *lat*$_1$, *time*$_1$, {A1}), (*lon*$_2$, *lat*$_2$, *time*$_2$, {A1}), ..., (*lon*$_4$, *lat*$_4$, *time*$_4$, {0}), ..., (*lon*$_7$, *lat*$_7$, *time*$_7$, {A2}), ..., (*lon*$_9$, *lat*$_9$, *time*$_9$, {A2, A3}), ..., (*lon*$_{14}$, *lat*$_{14}$, *time*$_{14}$, {A3}), 0 means there is no place to match this trajectory node. Third, as emphasised in Chapter 4, the temporal sequence must be considered in the process of stop episode detection, that is, all points in one cluster (reference spot or dense region) must be temporally closed. That is, the sequential order of trajectory nodes should be continuous, and a sudden increase between two points should not be permitted. If some trajectory nodes in a cluster satisfy the continuity of sequence, and the time duration of each of these points is greater than a user-specified threshold, these points are combined to form a stop episode. For example, Figure 5.3(a) shows there is a trajectory through two rectangular shaped places (left and right) where the numbers in circles represent the order of trajectory nodes and the blue solid circle c1 is the centroid of the left place. The trajectory nodes in the right rectangular place present some trajectory nodes between point 9 and point 30. Figure 5.3(b) shows that we need to identify possible stops first and then cluster those stops to identify stop episodes. Let us assume that p_1, p_2, p_3, p_6, p_7, p_9, p_{32}, p_{33}, p_{34} are possible stops as shown in Figure 5.3(b). Once potential stops are identified, then use the user-specified time duration to form stop episodes. In this particular example, let us set a time duration threshold to 30 minutes. Then, the duration of a set of continuous stops that is less than the user-provided time duration is not considered to be a stop episode (meaningful activity). Let us assume that time_duration(p_1, p_2, p_3) is 30 minutes, time_duration(p_6, p_7) is 45 minutes, time_duration(p_{32}, p_{33}, p_{34}) is

20 minutes, and time_duration(p_9) is 10 minutes. Then, (p_1, p_2, p_3) is a stop episode, and (p_6, p_7) is another stop episode as identified in Figure 5.3(c). In some cases, two stop episodes could be merged together to form one stop episode when they are spatially in the same place, and temporally close. In this particular example, two stop episodes $E_{1,3}$ and $E_{6,7}$ are in the same place, and the time duration of these two stop episodes are close enough (this is application-dependent and is less than a user-specified threshold). These two stop episodes could then be merged to represent one large stop episode $E_{1,7}$ as shown in Figure 5.3(d).

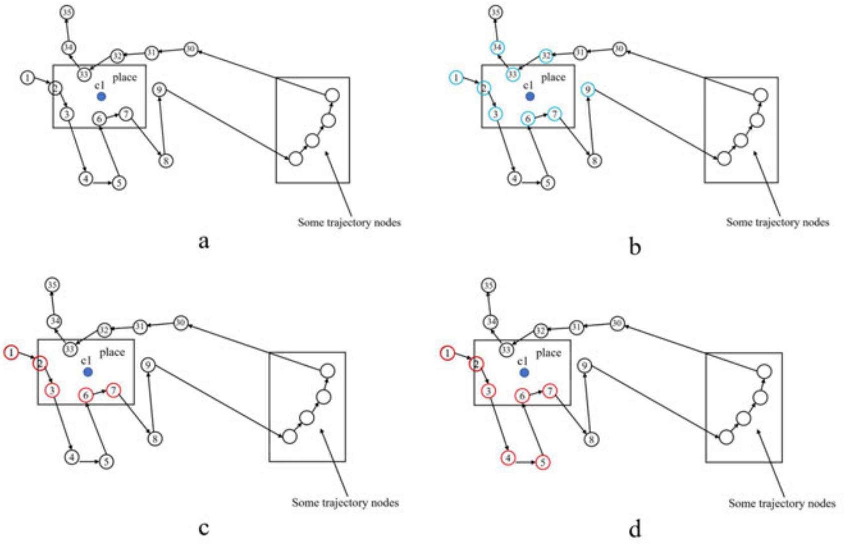


FIGURE 5.3: Identification of stop episodes.

Algorithm 3 illustrates a detailed procedure for stop episode detection.

5.2.3 Place Matching

In reality, there are two topological relationships among buildings: some places are spatially distinct and distant from each other or some places are spatially close. With the former, it is relatively easy to match each stop episode to a place, but with the latter one stop episode could be shared by a number of places and it is difficult to match the stop episode to a place. Figure 5.4 illustrates an example where a place A1 is distant from other places whilst places (A2 and A3), (A3 and A4), and (A5 and A6) are spatially close to each other. There is a trajectory passing through these places. Since A1 is distant from other places, it is relatively clear that stop episode 1 belongs to A1. Note that stop episode 2 (green points with red circles) is generated by A2, and is also part of stop episode 3, which is generated by A3. In this case, stop episode 2 is for A2 whilst stop episode 3 is for A3. When considering the sequential order and minimum time duration in clustering, the clusters are divided into different stop episodes when these clusters are spatially and temporally close. One example is when episode 3 completely includes episode 2 as shown in Figure 5.4. If all entries (stops) in

Algorithm 3 Stop episode detection

```
INPUT: A spatio-temporal trajectory $T$, ($\langle x_1, y_1, t_1 \rangle$, $\langle x_2, y_2, t_2 \rangle, \ldots, \langle x_m, y_m, t_m \rangle, \ldots, \langle x_n, y_n, t_n \rangle, \ldots, \langle x_k, y_k, t_k \rangle$), $k$ is length of
    trajectory, and semantic background information (.OSM file);
OUTPUT: A set of stop episodes $S$;
/* Divide all points into cluster points */
Extracts centroids for places;
Apply DBSCAN to get clusters C = $\{c_1, c_2, \ldots, c_k\}$ from $T$ and centroids
    for places;
/* stop episode finding */
for all $p \in c_j$ do
    if $p_m$ to $p_n$ is temporally sequential and $t(p_n) - t(p_m) > threshold$ then
        $p_m$, ..., $p_n$ is a stop episode;
        add $p_m$, ..., $p_n$ to $S$;
    end if
end for
/* combine stop episodes */
for all $se_j \in S$ do
    /* combine two continuous stop episodes*/
    if time_duration($se_j$, $se_{j+1}$) < $threshold$ and spatially close then $se_j$
    and $se_{j+1}$ are combined as a stop episode;
    end if
end for
```

a stop episode are shared by a few places, we will use HMM to match this stop episode to relevant places. For example, all entries in stop episode 4 are shared by place A3 and A4, stop episode 7 is shared by A7 and A8. Note that stop episode 5 and stop episode 6 share some entries (two grey points with red circles, where not one stop episode fully includes another stop episode). In this case, we will combine two stop episodes, and then use HMM to match this whole stop episode to the places.

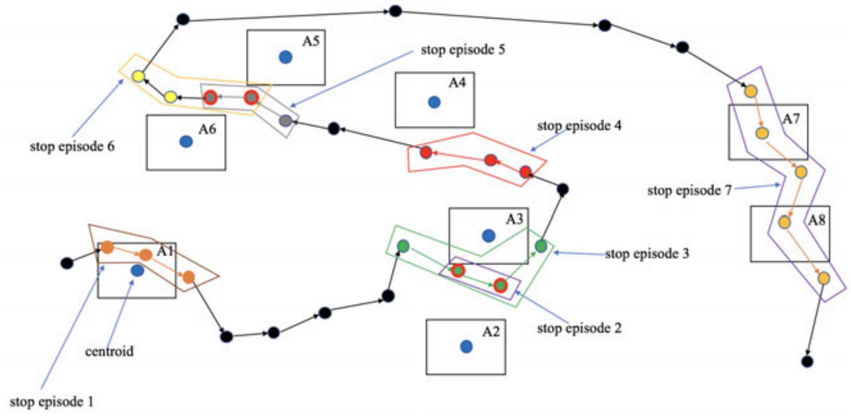


FIGURE 5.4: Illustration of place matching.

Hidden Markov Model for Place Matching

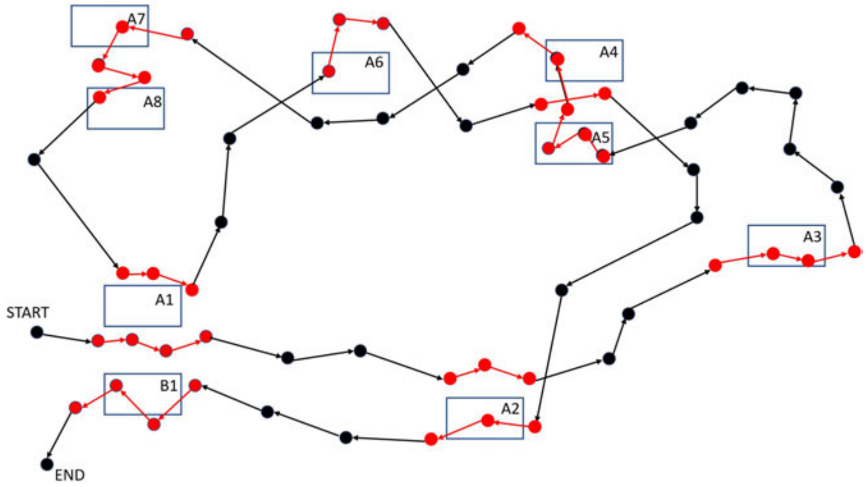


FIGURE 5.5: Illustration of place matching using HMM.

HMM (Rabiner, 1986) is used to match places in ambiguous and complex situations. HMM includes 3 main parts: initial probability; transition probability; and emission probability. Figure 5.5 shows how to use HMM for place matching. A series of connected red points (a trajectory segment) represents a unique stop episode. Initial probability is obtained by the place where the first stop episode belongs to. For example, in Figure 5.5, if the first stop episode belongs to A1, initial probability can be obtained as Table 5.1. If the first stop episode is ambiguous, that is, it belongs to A1 or B1, then these two buildings share the initial probability, 0.5 each.

TABLE 5.1: Initial probability for the example shown in Figure 5.5.

A1	B1	A2	A3	A4	A5	A6	A7	A8
1	0	0	0	0	0	0	0	0

TABLE 5.2: Transition probability for the example shown in Figure 5.5.

	A1	B1	A2	A3	A4	A5	A6	A7	A8
A1	0	0	0.5	0	0	0	0.5	0	0
B1	0	0	1	0	0	0	0	0	0
A2	0	0.5	0	0.5	0	0	0	0	0
A3	0	0	0	0	0.5	0.5	0	0	0
A4	0	0	0.5	0	0	0	0	0.25	0.25
A5	0	0	0.5	0	0	0	0	0.25	0.25
A6	0	0	0	0	0.5	0.5	0	0	0
A7	1	0	0	0	0	0	0	0	0
A8	1	0	0	0	0	0	0	0	0

Table 5.2 shows the result of transition probability. The first stop episode is ambiguous, it can be shown as (A1, B1), thus, all stop episodes can be represented by the places, like (A1, B1) $\rightarrow$ A2 $\rightarrow$ A3 $\rightarrow$ (A4, A5) $\rightarrow$ (A7, A8) $\rightarrow$ A1 $\rightarrow$ A6 $\rightarrow$ (A4, A5) $\rightarrow$ A2 $\rightarrow$ B1. There are 2 paths from A1 to other places, that is (A1, B1) $\rightarrow$ A2 and A1 $\rightarrow$ A6, thus we put A1 $\rightarrow$ A2 is 0.5, A1 $\rightarrow$ A6 is 0.5. A4 can be represented (A4, A5) $\rightarrow$ (A7, A8) and (A4, A5) $\rightarrow$ A2, due to A7 and A8 share the stop episode, thus the probability for A7 and A8 are 1/2 * 1/2 = 1/4 and A4 $\rightarrow$ A2 is 1/2.

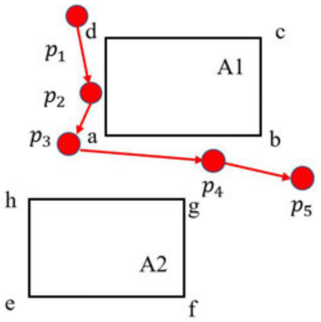


FIGURE 5.6: An example illustrating no intersection in place matching: two rectangular places A1 and A2 with E_{p_1,p_5}; $abcd$ are four corners of A1 whilst $efgh$ are four corners of A2.

The last part is how to compute emission probability. There are two situations we need to consider and they are whether the stop episodes intersect with places or not. If not, like an example shown in Figure 5.6, the stop episode (p_1, p_2, p_3, p_4, p_5) could belong to either A1 or A2. This stop episode does not intersect with A1 and A2, thus, the emission probability for this stop episode belonging to A1 can be computed as follows:

$$total\_time\_duration = time\_duration(p_1, p_5), \tag{5.1}$$

$$\begin{aligned} total\_distance =&dist(\overline{p_1p_2}, \overline{ad}) + dist(\overline{p_2p_3}, \overline{ad}) + \\ &dist(\overline{p_3p_4}, \overline{ab}) + dist(\overline{p_4p_5}, \overline{ab}), \end{aligned} \tag{5.2}$$

$$\begin{aligned} P(A1) =&\frac{time\_duration(p_1, p_2)}{total\_time\_duration} * (1 - \frac{dist(\overline{p_1p_2}, \overline{ad})}{total\_distance}) + \\ &\frac{time\_duration(p_2, p_3)}{total\_time\_duration} * (1 - \frac{dist(\overline{p_2p_3}, \overline{ad})}{total\_distance}) + \\ &\frac{time\_duration(p_3, p_4)}{total\_time\_duration} * (1 - \frac{dist(\overline{p_3p_4}, \overline{ab})}{total\_distance}) + \\ &\frac{time\_duration(p_4, p_5)}{total\_time\_duration} * (1 - \frac{dist(\overline{p_4p_5}, \overline{ab})}{total\_distance}), \end{aligned} \tag{5.3}$$

Where the total time duration ($total\_time\_duration$) is computed by the time duration of a given stop episode, the total distance ($total\_distance$) is computed by the sum of distances between two line segments (a line segment with two successive nodes in the stop episode and the nearest line segment of the rectangular place), $dist(.,.)$ computes a distance between two given points, and $P(A1)$ computes an emission probability for A1.

The total distance between a stop episode and a place is calculated in Equation 5.2. Note that we choose the nearest edge of a rectangular place for each segment of a stop episode to compute the distance. For example, we compute the distance between $\overline{p_1p_2}$ and $\overline{ad}$ instead of $\overline{p_1p_2}$ and $\overline{ab}$. Based on Equation 5.1 and 5.2, the emission probability between a stop episode and a place is computed in Equation 5.3. In the process of place matching, the time duration and the distance between a stop episode and a place are two key factors because: (1) time duration of each segment (two consecutive entries) in a stop episode is different. If one segment spends a longer time, then it means the object stays longer which implies this segment is more important than other segments from the temporal aspect; and (2) if one segment in a stop episode is spatially closer to a place than another segment, this segment is more important from the spatial aspect.

In reality, an object usually enters into a building with a certain purpose such as shopping or working. Figure 5.7 shows that one stop episode intersects with places. Figure 5.7 (a) shows there is a stop episode (p_1, p_2, p_3, ,p_4, p_5, p_6) passing through places A1 and A2. Equation 5.4 illustrates how to compute an emission probability for the stop episode (p_1, p_2, p_3, p_4, p_5, p_6) to match A1. First, we first find entries in a stop episode which is relevant to place A1, that is, finding the last stop before entering into place A1 (p_4) and the first stop coming out from A1 (p_6). We compute a time duration between them, which indicates how long the object stays in A1. If it is long, it means a segment $\overline{p_4p_6}$ is important from the temporal aspect. Spatially, each stop episode has a centroid, a radius, a starting time, and a duration. The centroid is calculated by the average spatial coordinate of all the stops in the episode. After the stop centroid is known, the stop radius is calculated as

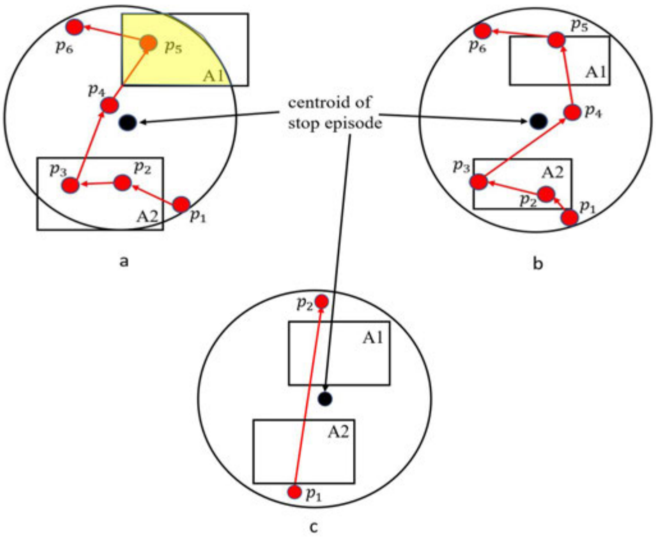


FIGURE 5.7: An example illustrating intersection in place matching.

the distance from the centroid to the most distant stop in the stop episode. As shown in Figure 5.7, a stop episode circle can be computed by a stop centroid and a stop radius. The yellow part of the circle which overlaps place A1 is calculated by the area of the stop episode circle $(Area(stopcircle))$ $\cap$ the area of place A1 $(Area(A1))$ as shown in Equation 5.4. In Equation 5.4, w_1 and w_2 are weights and the sum of these two weights is 0.5. Note that considering the time duration is also of importance. Figure 5.7 (b) illustrates that the stop circle completely includes A1 and A2. In this example, it is necessary to consider both the spatial relationship and the temporal dimension. Finally, Figure 5.7 (c) illustrates an extreme situation when there are only two entries in a stop episode and its stop episode circle completely includes A1 and A2. In this case, we assign 0.5 to A1 and A2 to make them equally contribute to the stop episode. After computing initial probability, transition probability and emission probability, we use Viterbi algorithm (Viterbi, 1967) to find the most possible sequence of place visitations for each stop episode.

$$\begin{aligned} P(A1) = 0.5 + w_1 * \frac{time\_duration(p_4, p_6)}{time\_duration(p_1, p_6)} + \\ w_2 * \frac{Area(stopcircle) \cap Area(A1)}{Area(A1)}, \end{aligned} \tag{5.4}$$

where w_1 and w_2 are weights and the sum of them is equal to 0.5.

5.2.4 Period Detection

Past studies in the reference spot approach (Li et al., 2010a; Li et al., 2012; Zhang et al., 2018) use the combination of Fourier transform and autocorrelation in order to detect regular periods. One limitation of Fourier transform is that it requires evenly spaced (regular) time series data as input. However, unevenly spaced data are very often in the real world due to a variety of reasons such as limitations of instruments and errors in devices. Therefore, these traditional period detection approaches cannot be directly used for irregular spatio-temporal trajectories under study. In this chapter, we utilise Lomb-Scargle periodogram (Lomb, 1976; Scargle, 1982) in order to handle unequally sampled/recorded and irregular trajectories.

Lomb (1976) proposed an approach to find periods in unevenly spaced data where he used least squares fits to sinusoidal curves. Scargle (1982) extended Lomb's work by defining Lomb-Scargle periodogram. Ruf (1999) was one of the first to employ Lomb-Scargle periodogram for the analysis of biological data. He used this technique to detect a circadian rhythm with a period of 24 hours for alpine marmot, based on telemetric temperature data. Glynn et al. (2006) used Lomb-Scargle periodogram to detect significant periodic gene expression patterns. They proved that Lomb-Scargle periodogram is an effective approach to finding periodic gene expression profiles in microarray data, especially when data are sampled at arbitrary time points or when missing data exist at a significant proportion. Van Dongen et al. (1999) successfully analysed unevenly spaced time series data of human oral temperatures with a period of 24 hours. Bohn et al. (2003) applied Lomb-Scargle periodogram for periodogram estimation.

For time series data, considering N observations where x_j was taken at time t_j for $j = 1, 2, \ldots, N$, Lomb-Scargle periodogram is calculated by Equation 5.5:

$$\begin{aligned} P_{LS}(f) = \frac{1}{2\sigma^2}\{ & \frac{[\sum_{j=1}^{N}((x_j - \overline{x})\cos(2\pi f(t_j - \tau))]^2}{\sum_{j=1}^{N}\cos^2(2\pi f(t_j - \tau))} + \\ & \frac{[\sum_{j=1}^{N}(x_j - \overline{x})\sin(2\pi f(t_j - \tau))]^2}{\sum_{j=1}^{N}\sin^2(2\pi f(t_j - \tau))}\}, \end{aligned} \tag{5.5}$$

$$\overline{x} = \frac{1}{N}\sum_{j=1}^{N} x_j, \tag{5.6}$$

$$\sigma^2 = \frac{1}{N-1}\sum_{j=1}^{N}(x_j - \overline{x})^2, \tag{5.7}$$

where $\bar{x}$ and σ^2 are the mean and the variance of the time series data as shown in Equation 5.6 and 5.7, and τ is specified for each f to ensure time-shift invariance caused by unevenly sampled (irregular) data, which is shown in Equation 5.8.

$$\tan(2(2\pi f)\tau) = \frac{\sum_{j=1}^{N}\sin(2(2\pi f)t_j)}{\sum_{j=1}^{N}\cos(2(2\pi f)t_j)}. \tag{5.8}$$

The false alarm probability in Lomb-Scargle periodogram was shown by Scargle (1982), and the false alarm probability of peaks P_{max} in the periodogram caused by a chance noise fluctuation can be calculated by Equation 5.9.

$$Pr(p_{max}) = 1 - [1 - exp(-p_{max})]^N. \tag{5.9}$$

From the distribution in Equation 5.9, to find a power level z where a peak must exceed to reach the statistical significance at a given error probability α (for example 0.01, 0.05) is computed by Equation 5.10

$$z = -\ln[1 - (1 - \alpha)^{\frac{1}{N}}]. \tag{5.10}$$

Please refer to Lomb-Scargle periodogram (Lomb, 1976; Scargle, 1982) for more details.

5.3 Experimental Results

This section includes explanation of two real datasets that we use in this study, evaluation of efficiency and effectiveness for existing approaches and our method. Note that the effectiveness of reference spots and the effectiveness of periods are two key steps. Effective periods can be obtained by high quality reference spots; thus the effectiveness of reference spots should be explained before the evaluation of effectiveness of periods. Finally, we show the effectiveness of periodic patterns in this section.

5.3.1 Datasets

Two real datasets (referred to as Dataset 1 and Dataset 2) are used for experimental studies in this chapter. Due to the nonexistence of ground-truth dataset for PPM evaluation, the first synthetic real-world GPS dataset was intentionally collected by authors from 20/9/2017 to 20/10/2017 in Cairns, Australia. This generated a ground-truth benchmark dataset in order to validate and evaluate our proposed approach against past studies. There are two thing to consider in this dataset. First, to measure the effectiveness of proposed algorithm for PPM, authors intentionally visited a range of places regularly so that the dataset would not record the author's normal routines. Second, authors periodically visited some places with regular time intervals in order to generate some periodic patterns. Dataset 1 is shown in Figure 5.8(a). The second GPS dataset is from Geolife [2], which was collected by Microsoft Research Asia. The Geolife project was compiled by 182 users over a five-year period (from 4/2007 to 8/2012) in Beijing, China. The dataset collected users' outdoor movements including daily life, for example, their regular travels between home and work place, entertainment, sports, study, and shopping activities. We utilised one of the users from Geolife for our study which was recorded

[2] https://privamov.github.io/accio/reference/datasets/

from 24/10/2008 to 23/11/2008 (this is referred to as Dataset 2). Figure 5.8(b) displays Dataset 2. Table 5.3 summarises the main features of these two datasets under study, which are both GPS data. They are irregular trajectories and the time period is approximately one month.

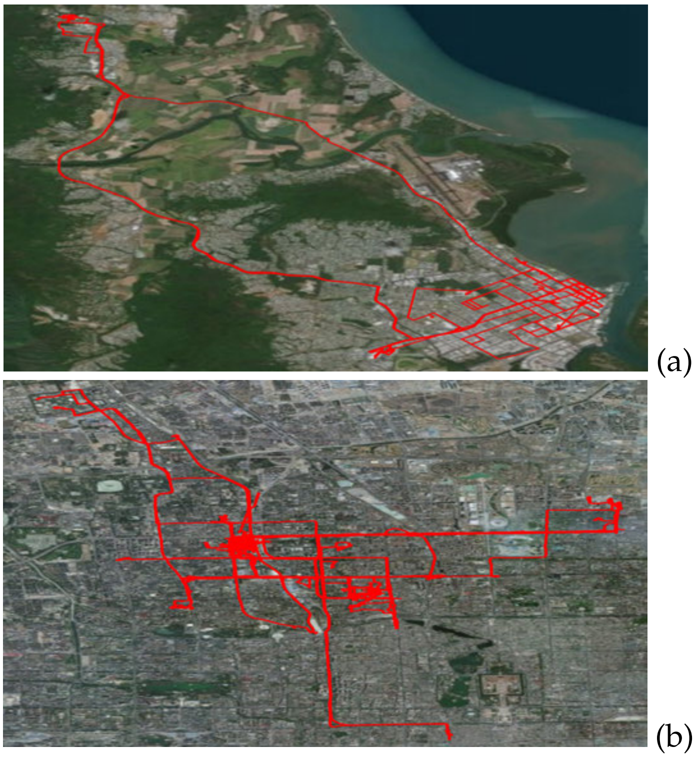


FIGURE 5.8: Visualisations of Dataset 1 and Dataset 2.

TABLE 5.3: A summary of two datasets under study.

	Type	Time	Time
		Period	Interval
Dataset 1	GPS	1 month	irregular
Dataset 2	GPS	1 month	irregular

5.3.2 Trajectory Interpolation

Past reference spot approaches (Li et al., 2010a; Li et al., 2012) need an interpolation preprocessing step to transform irregular spatio-temporal trajectories into regular ones for period detection. To compare the effectiveness and efficiency of our method against these traditional approaches, we apply the most common linear interpolation (Rhee et al., 2011; Li et al., 2010b) to the existing approaches to conduct subsequent comparative experiments.

Efficiency for Linear Interpolation

Figure 5.9(a) and (b) display the running time analysis using linear interpolation with different time intervals (10, 30, 60 , 90, 120 seconds) for Dataset 1 and Dataset 2. As the time interval increases, the running time greatly decreases. It becomes very time-consuming when 10 seconds is used. Even though we use a long interval, 120 seconds as an interpolation interval, the running time requires almost 60 seconds and 8 seconds on linear interpolation for Dataset 1 and Dataset 2.

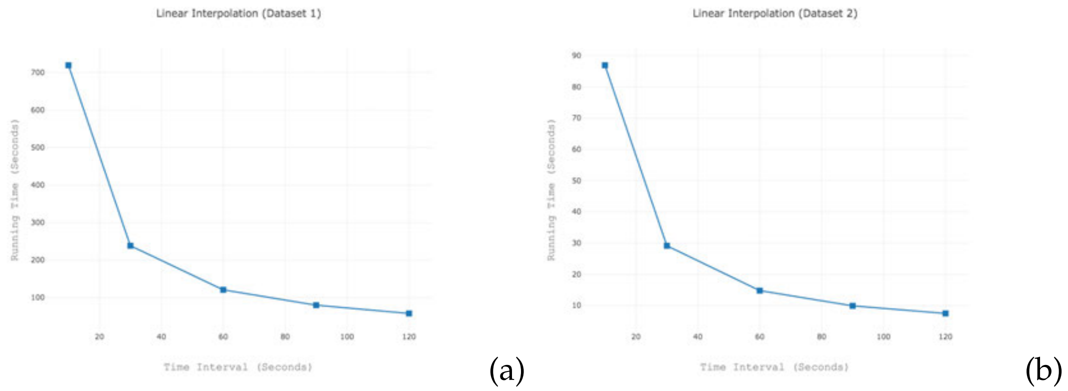


FIGURE 5.9: Efficiency analysis with different time intervals in linear interpolation: (a) Dataset 1; (b) Dataset 2.

Effectiveness for Linear Interpolation

In this section, we show the number of references spots obtained from Periodica and Traclus (ST) after applying linear interpolation with different time intervals, 10, 60 and 120 seconds.

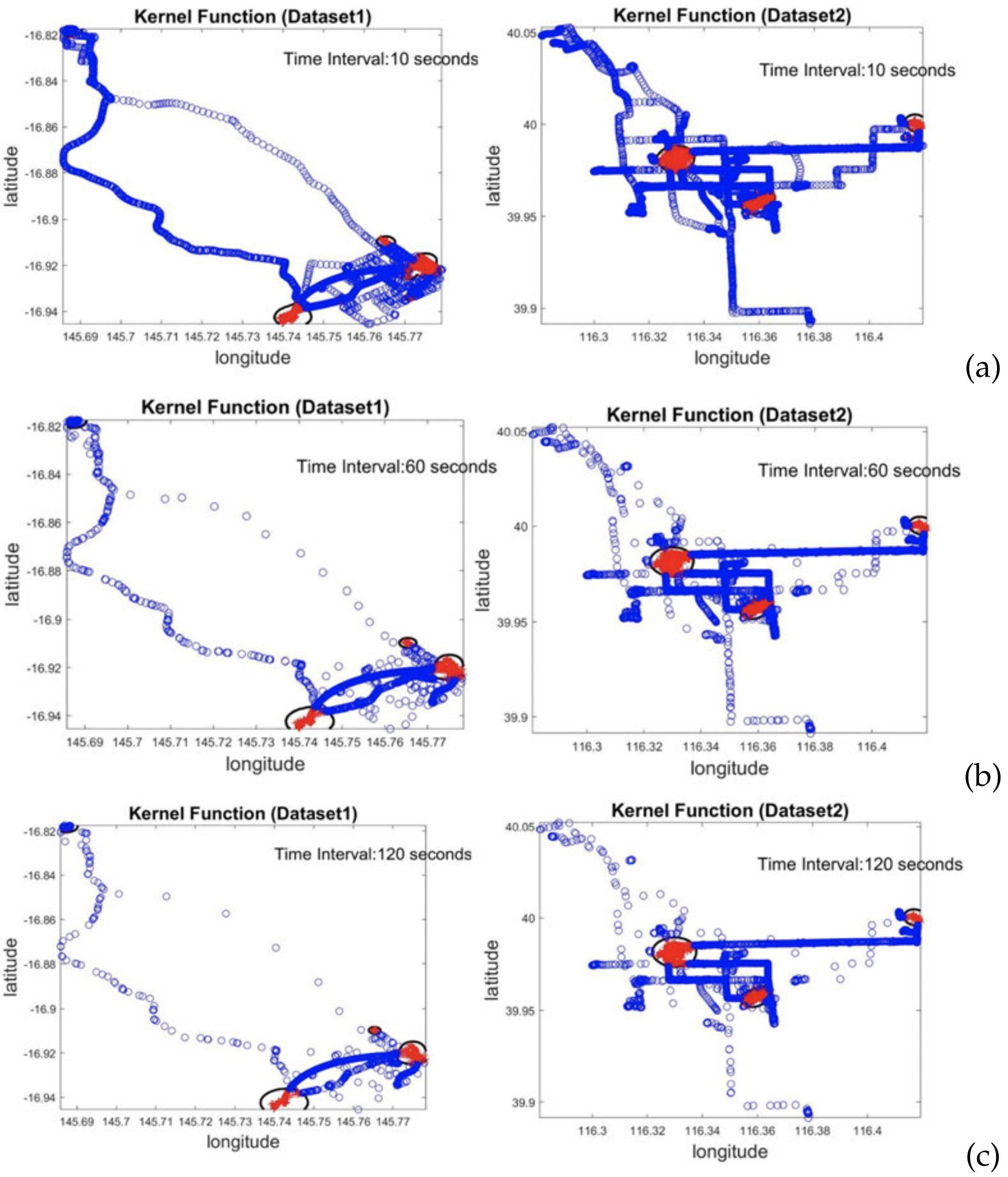


FIGURE 5.10: Reference spots with Periodica for Dataset 1 and Dataset 2: (a) 10 seconds interpolation; (b) 60 seconds interpolation; (c) 120 seconds interpolation.

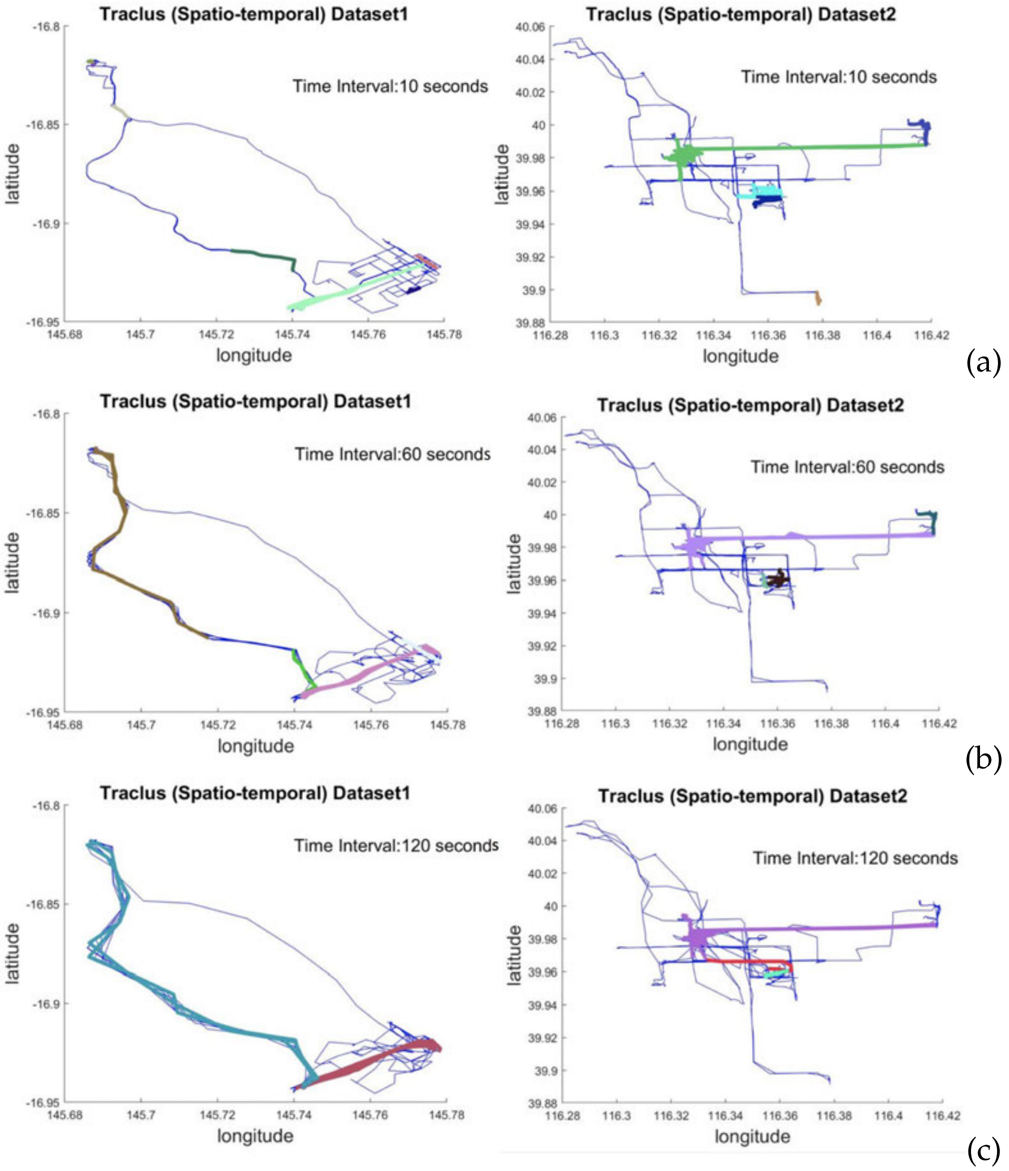


FIGURE 5.11: Reference spots with Traclus (ST) for Dataset 1 and Dataset 2: (a) 10 seconds interpolation; (b) 60 seconds interpolation; (c) 120 seconds interpolation.

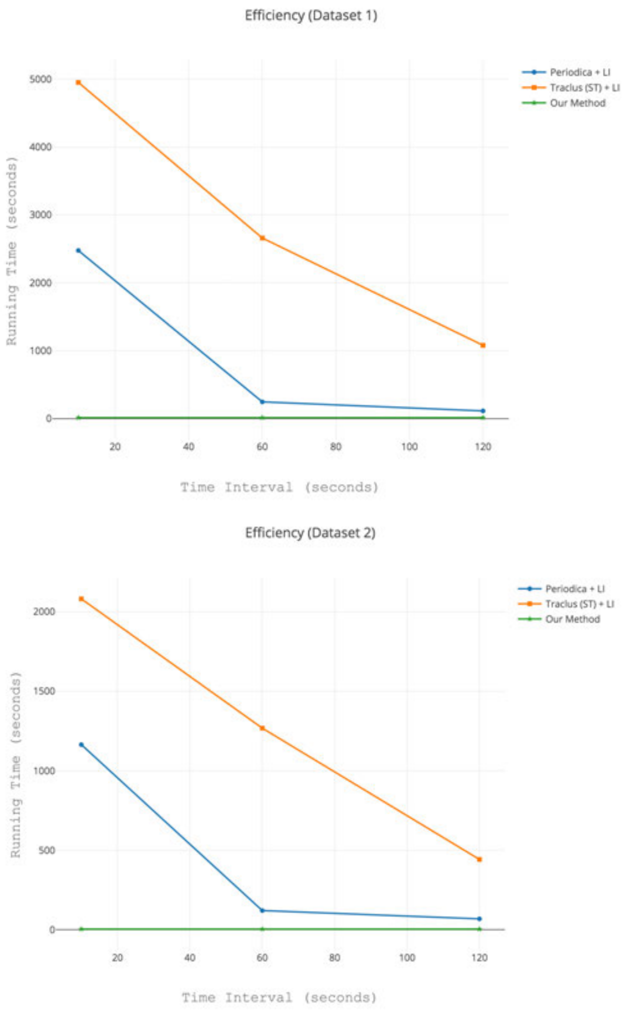
TABLE 5.4: The number of reference spots for Dataset 1 and Dataset 2.

Approach	Dataset 1	Dataset 2
Periodica (10 seconds)	6	4
Periodica (60 seconds)	4	3
Periodica (120 seconds)	4	3
Traclus (ST) (10 seconds)	7	5
Traclus (ST) (60 seconds)	4	4
Traclus (ST) (120 seconds)	2	3
Our approach	10	10

Figure 5.10 and Figure 5.11 visualise reference spots obtained from Periodica and Traclus (ST). The main difference between them is Traclus (ST) can find some dense paths between places whilst Periodica tends to find some compact clusters. Table 5.4 shows the number of reference spots from two existing algorithms. Regardless of whether Periodica or Traclus (ST) is in place, they are both sensitive to the number of trajectory nodes. On the one hand, it will be relatively fast when a large interpolation interval is used. However, this large time interval fails to model local variations and details, and eventually misses some reference spots. On the other hand, when a small interval is in place, it becomes time-consuming, but it captures more details and variations, and thus produces more reference spots than the use of large interpolation interval. There is a trade-off between efficiency and effectiveness with the use of interpolation interval. Finding the best interpolation interval is data-dependent and requires several trial-and-error steps which are difficult and time-consuming.

5.3.3 Efficiency

Figure 5.12 displays the efficiency analysis of three approaches for Dataset 1 and Dataset 2. Although Periodica exhibits better efficiency than Traclus (ST), Periodica and Traclus (ST) are both inefficient, especially when 10 seconds is used as the time interval. It becomes computationally inefficient. Even when we use 120 seconds as the time interval, Periodica still spends 114.624196 seconds and 68.748305 seconds for Dataset 1 and Dataset 2, respectively. Note that, our method does not need an interpolation step, and it only requires 15.2869 seconds and 4.549142 seconds for Dataset 1 and Dataset 2, respectively, which is much faster than Periodica and Traclus (ST), even with the use of 120 seconds as the time interval.



(a)

(b)

FIGURE 5.12: Comparison of the efficiency of three methods.

5.3.4 Effectiveness of Reference Spots

As previously mentioned in Chapter 3, a reference spot is a dense region where the moving object frequently visits. The high quality of reference spot is very important to find useful periodic pattern. In this section, we evaluate the performance of three algorithms for place extraction. We compare them in five ways to measure the effectiveness of reference spots: (1) number of reference spots; (2) spatial compactness; (3) temporal compactness; (4) spatio-temporal compactness; (5) semantic accuracy.

Number of Reference Spots

Table 5.4 shows that our method can find more reference spots than Periodica and Traclus (ST), even when 10 seconds as the time interval is used to find the maximum number of reference spots at the expense of efficiency. Our approach detects the references spots detected by Periodica and Traclus (ST) and additionally finds local clusters. Therefore, our approach is less vulnerable to false positive than Periodica and Traclus (ST). One other interesting finding to note is that as the time interval decreases, Traclus (ST) obtains more reference spots than Periodica. When more local details and variations are captured and modelled with the use of small interpolation intervals, it becomes important to consider the temporal and sequential aspect in order to take sequential variations into account in reference spot detection. Traclus (ST) tends to generate more reference spots when a smaller interpolation interval is used.

Spatial Compactness

5.13 shows 10 single-level semantic places that are extracted from the data owner's activities and daily life. The green numbers and arrows indicate *i*-th place. Table 5.5 lists 10 single-level semantic places of Dataset 1 with their place names. These are major semantic places where the data owner visits and spends most time. These places include the data owner's home, university buildings for studying, shopping malls for shopping activities and dining, gym for exercise, Holiday Inn for a part-time job, and RLS for learning activities. Trajectory nodes in the same semantic place should be spatially concentrated and close, thus a reference spot with good quality should exhibit a high spatial compactness and closeness value. The spatial compactness of a semantic place is calculated by the average distance between the centroid of place and all points in one reference spot which are close to or covering the place (Legány et al., 2006), and the spatial compactness of an approach is the sum of spatial compactness of all reference spots identified by the approach, which means the smaller average distance, the higher spatial compactness.

We use a 10 second interval for the interpolation process for Periodica and Traclus (ST) in order for them to produce quality reference spots. We test spatial compactness with Dataset 1 since we have the ground-truth for the dataset. Figure 5.14 shows the superior performance of our approach. In the figure, x-axis represents i-th semantic place and y-axis shows its corresponding average distance. Higher average distance exhibits lower spatial compactness. Most of reference spots obtained from our method exhibit higher spatial compactness (lower average distance) than Periodica and Traclus (ST). Not surprisingly, Traclus (ST) exhibits the worst performance in spatial compactness since it not only optimises spatial compactness but also considers temporal sequential compactness. Note that Periodica utilises Kernel function considering spatial densities whilst ignoring the temporal sequential aspect to find reference spots, and thus it performs better than Traclus (ST) for all reference spots, and even for reference spots 2, 3 and 6 where Periodica obtains lower average distance values than our method. However, in general even though our approach considers three essential aspects: spatial, temporal and aspatial semantic information, it outperforms Periodica. The average spatial compactness of Periodica for Dataset 1 is 0.00165, and for Traclus (ST) is 0.0046, whilst that of our method is 0.00043, which demonstrates the superiority of our approach in this spatial compactness.

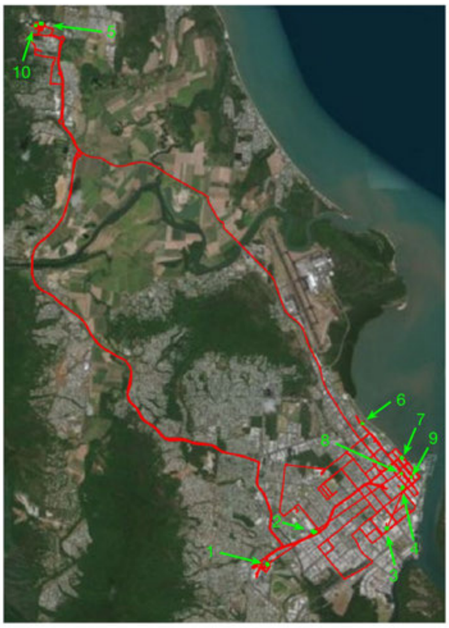


FIGURE 5.13: Semantic places in Dataset 1.

TABLE 5.5: Annotations of semantic places.

Number	Semantic Places
1	Author's Home
2	DFO (Direct Factory outlet Shopping Mall)
3	Transportation Office
4	Rusty Market (A Sunday Market)
5	Post Graduate Center (University)
6	Holiday Inn
7	RLS (A toastmaster Club)
8	Gym
9	WoolWorth (Supermarket)
10	Vice-Chancellor Building (University)

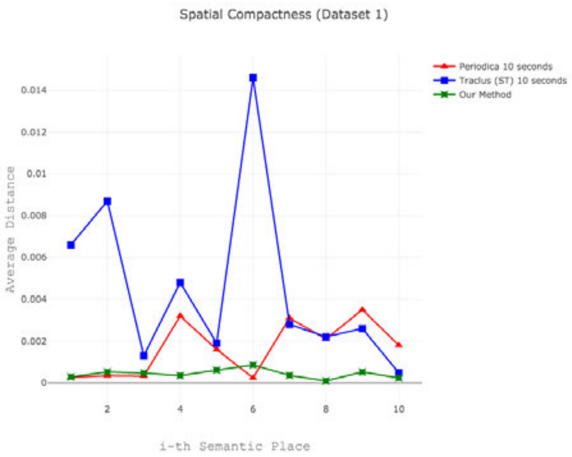
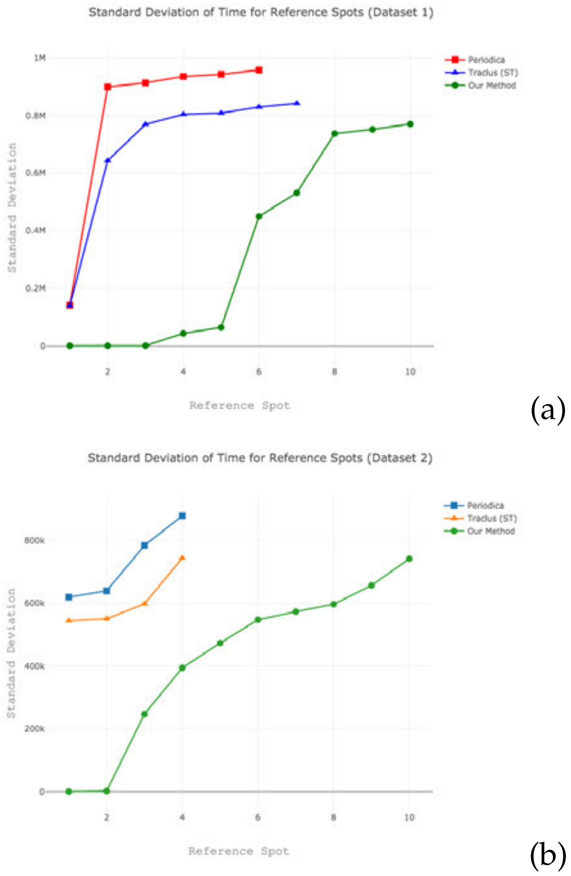


FIGURE 5.14: Measure of spatial compactness with the three methods for Dataset 1 (average spatial compactness: 0.00165 for Periodica; 0.0046 for Traclus (ST); 0.00043 for our method).

Temporal Compactness



(a)

(b)

FIGURE 5.15: Measure of temporal compactness with the three methods for Dataset 1 and Dataset 2 (average temporal compactness: 798420 for Dataset 1 and 730956 for Dataset 2 for Periodica; 690940 for Dataset 1 and 609780 for Dataset 2 for Traclus (ST); 334990 for Dataset 1 and 423590 for Dataset 2 for our approach.)

Reference spots should be spatially concentrated and also temporally aggregated, therefore in this section we measure the temporal compactness. There are some measures to represent spread, but standard deviation is one popular and widely used approach to measure spread. Figure 5.15 shows the measure of temporal compactness with the three methods under study for Dataset 1 and Dataset 2. Note that we sort the standard deviation values in an ascending order for better comparison and visualisation. In this figure, x-axis represents i-th reference spot for each method whilst y-axis shows a corresponding standard deviation value for each reference spot. Higher standard deviation exhibits lower temporal compactness. Not surprisingly, most of the reference spots in Periodica result in higher standard deviation values than those in Traclus (ST) and in our method. That is, Periodica performs the worst since it does not consider the temporal sequential aspect when detecting reference spots. Thus temporally distant points could be grouped into the same cluster which increases the temporal spread within clusters. This implies that Periodica could generate false positive reference spots. Both Traclus (ST) and our approach consider the sequential aspect producing better temporal compactness. However, our approach outperforms Traclus (ST) by a great margin as shown in Figure 5.15. Figure 5.15 (a) displays a temporal compactness graph for Dataset 1 where the average standard deviation values (temporal compactness) are: 798420 for Periodica, 690940 for Traclus (ST) and 334990 for our approach. Note that the standard deviations of half the number of reference spots in our method are lower than those of reference spots in Periodica and Traclus (ST). The largest standard deviation for a reference spot in our method is still lower than most of those reference spots in Periodica and Traclus (ST). On the other hand, Figure 5.15 (b) shows a temporal compactness graph for Dataset 2 where the average standard deviation values (temporal compactness) are: 730956 for Periodica, 609780 for Traclus (ST) and 423590 for our approach. This clearly demonstrates the superior performance of our approach over Periodica and Traclus (ST) with regard to temporal compactness.

Spatio-temporal Compactness

Spatio-temporal compactness combines both spatial compactness and temporal compactness. Since both space and time have different scales, we first normalise each compactness into the same scale ([0,1]) using the min-max normalisation in order to avoid the unnecessary scale effect (Bermingham and Lee, 2017). In this study, space and time are equally important and the same weight (0.5) is assigned to each of these compactness values, where the sum of weight values is equal to one. Thus, spatio-temporal compactness for a reference spot is calculated by the summation of 0.5 × min-max normalised corresponding spatial compactness and 0.5 × min-max normalised corresponding temporal compactness. Thus, spatio-temporal compactness for a reference spot can be calculated by Equation 5.11. STC is spatio-temporal compactness, SC is spatial compactness and TC is temporal compactness.

$$STC = 0.5 \times normal(SC) + 0.5 \times normal(TC) \qquad (5.11)$$

Figure 5.16 displays the spatio-temporal compactness values of the three methods. Our method obtains the lowest value for each semantic place when compared to Periodica and Traclus (ST). One thing to note is that unlike the spatial compactness performance, Traclus (ST) exhibits better performance than Periodica in spatio-temporal compactness for each place. This is because Traclus (ST) considers both spatial and temporal aspects. Furthermore, our approach outperforms Traclus (ST) in this compactness even though both approaches consider two spatial and temporal aspects together. The average spatio-temporal compact values for Dataset 1 are: 0.8804 for Periodica, 0.7828 for Traclus (ST), and 0.3497 for our approach.

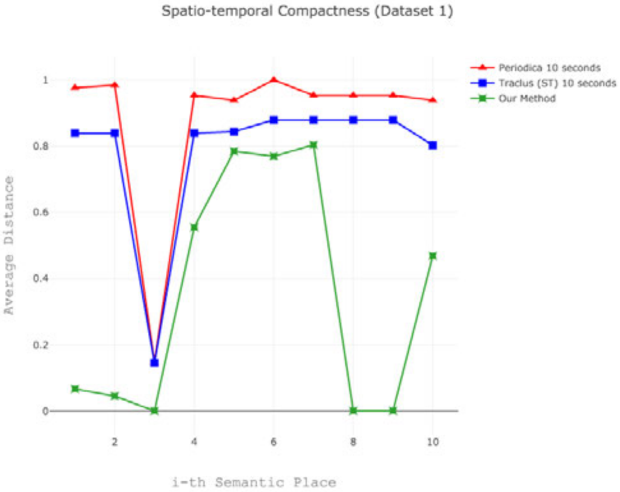
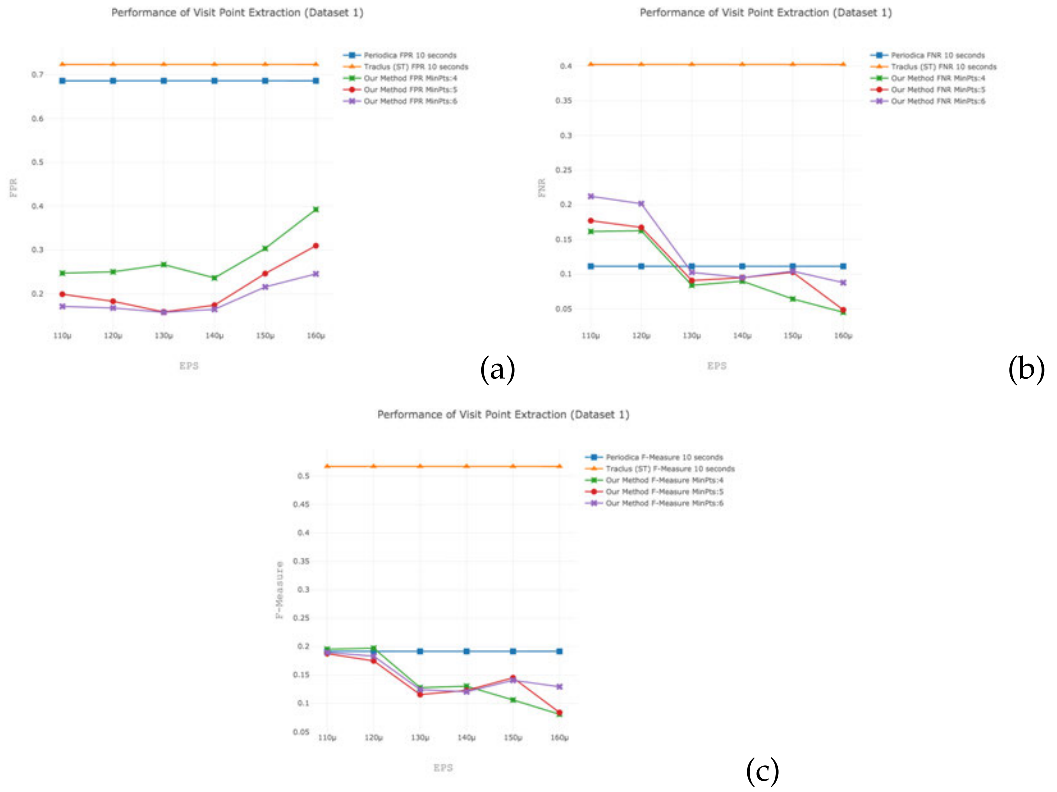


FIGURE 5.16: Measure of spatio-temporal compactness with the three methods for Dataset 1 (average spatio-temporal compactness: 0.8804 for Periodica; 0.7828 for Traclus (ST); and 0.3497 for our approach).

Semantic Accuracy

Since Dataset 1 has a set of semantic true positive and negative places, we measure semantic accuracy through false positive and false negative ratio. False positives are ones that algorithms report as stops and visiting places but are moving rather than actually visiting semantic places. False negatives are ones that algorithms classify as moving but are actually visiting semantic places. We set two main parameters of DBSCAN, *eps* and *minpts*, to reduce both false positive and false negative as far as possible. False Positive Rate (FPR) is the ratio between the number of negative instances incorrectly classified as positive and the total number of actual negative instances whilst False Negative Rate (FNR) is the ratio between the number of positive instances incorrectly classified as negative and the total number of actual positive instances. F-measure is the harmonic mean of FNR and FPR (Lv et al., 2016) which can be calculated by Equation 5.12. For Kernel function, we set p to 15% as recommended (Li et al., 2010a; Li et al., 2012) since most parts of reference spots are detected with this value. For Traclus (ST), as suggested by the authors, we use silhouette coefficient (Rousseeuw, 1987) to generate quality trajectory clusters.

$$F - measure = \frac{2 \times FPR \times FNR}{FPR + FNR}. \tag{5.12}$$



(a)

(b)

(c)

FIGURE 5.17: Performance comparison of the three methods for place extraction: (a) FPR graph; (b) FNR graph; (c) F-measure graph.

Figure 5.17(a), (b) and (c) show FPR, FNR and F-measure of the three algorithms for Dataset 1, respectively. For Periodica and Traclus (ST), we use a 10 seconds interpolation internal to compute FPR, FNR and F-measure since it produces more and better references spots for these methods than larger interpolation time intervals. For our method, we apply different eps and $minpts$ values to observe different performance behaviors of our approach. In Figure 5.17, x axis represents eps for our approach whilst y axis shows FPR, FNR and F-Measure. Note that Periodica and Traclus (ST) exhibit a constant performance behavior across different eps values. In Figure 5.17 (a), we compare Periodica and Traclus (ST) with our approach against FPR. Although Traclus (ST) obtains more reference spots than Periodica, Periodica exhibits a better performance on FPR than Traclus (ST). Not surprisingly, our approach outperforms Periodica and Traclus (ST) with all eps and $minpts$ values tested. Figure 5.17(b) shows FNR for the three methods. Traclus (ST) exhibits an extremely higher FNR than those of

Periodica and our method. Noted that, Periodica obtains a lower FNR for certain *eps* values (0.00011 and 0.00012) than our method, but in general our approach exhibits a better performance with regard to FNR. Figure 5.17(c) displays F-Measure for the three approaches. Traclus (ST) exhibits a higher F-Measure than Periodica and our method. Note that, Periodica shows a better performance in F-Measure than our approach when we use 0.00011 or 0.00012 for *eps* and 4 for *minpts*. Figure 5.17 shows the influence of two main parameters *eps* and *minpts* with our method on FPR, FNR and F-measure. First, it is clear to see that our method shows a better performance on FPR, FNR and F-measure than Periodica and Traclus (ST). The best overall performance on FPR, FNR and F-measure is obtained by *eps* = 0.00013, *minpts* = 5. Second, setting *eps* to a high value results in a high FPR since it tends to produce large clusters capturing temporally long movements whilst setting *eps* to a small value ends up with a high FNR since it tends to generate small clusters modelling only temporally short movements (thus temporally long movements are considered to be false negatives).

5.3.5 Effectiveness of Periods

Due to the known ground-truth for Dataset 1, in this section, we evaluate the effectiveness of obtained periods by Lomb-Scargle periodogram and Fourier transform and autocorrelation with regard to the ground-truth real periods. Our aim is to prove that although Lomb-Scargle periodogram does not need interpolation when compared to Fourier transform and autocorrelation, it outperforms traditional Fourier transform and autocorrelation for period detection. In order to compare the effectiveness of period detection between Lomb-Scargle periodogram used in our approach and the combination of FT&Auto used in (Li et al., 2010a; Li et al., 2012), we use the same set of reference spots detected by our approach. For the FT&Auto approach, we generate a binary sequence for each reference spot, with 1 indicating the object is in the reference spot whilst 0 indicating it is out of the reference spot. As stressed before, the Fourier transform and autocorrelation approach only accepts regular time interval as input, thus, we make a binary sequence with a regular time interval for comparison. That is, in Figure 5.18, a trajectory starts from 9am, and then the object stays in place 1 for 30 minutes, then stays in place 2 for 30 minutes, finally, the trajectory ends at 11:30am. After finding place 1 and place 2 as reference spots by our method, a binary sequence is generated as 011011110. Next, we represent the trajectory and also to make it regular. Note that when we use 10 minutes as a time interval, then the binary sequence becomes 0001111001111000. This binary sequence can be used for the Fourier transform and autocorrelation approach to detect regular periods.

Figure 5.19(a) displays obtained reference spots from our method for Dataset 1 (indicated by cluster numbers and arrows). The figure clearly shows that our method effectively finds reference spots for semantic places that are shown in Figure 5.13. Table 5.6 displays a list of places with non-zero periods for Dataset 1. Numbers for references spots in Table 5.6 and Figure 5.19(a) are reference spot identification numbers shown in Table 5.5. We compare real periods with obtained periods from Lomb-Scargle periodogram (irregular) and Fourier transform with autocorrelation. As in the previous experiments, we use 10, 60 and 120 seconds to interpolate binary sequences. Table 5.6 shows obtained periods from Lomb-Scargle periodogram, and Fourier transform and autocorrelation. Fourier transform and autocorrelation produces the same set of periods with 10 and 60 seconds time intervals, which is close to the ground-truth values. However, its performance deteriorates with 120 seconds, and it fails to produce non-zero periods for some places with 120 seconds as shown in Table 5.6. Lomb-Scargle periodogram performs slightly worse than Fourier transform and autocorrelation with 10 and 60 second intervals, but it outperforms Fourier transform and autocorrelation with 120 seconds. Note that, In Table 5.7, Fourier transform and autocorrelation requires more time as the interpolation interval becomes smaller, and the 10 seconds interval requires over 14 times more time than the 120 seconds interval, and the 60 seconds interval requires around two times more than 120 seconds. Noticeably, our approach requires around 40 times less time than Fourier transform and autocorrelation with a 120 second interval but outperforms it to a great degree. Also our approach produces near similar results to Fourier transform and autocorrelation with 60 seconds and with around 80 times less time.

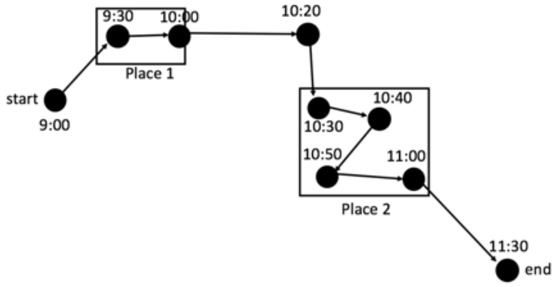


FIGURE 5.18: An example illustrating how to make a binary sequence regular.

TABLE 5.6: Period detection for Dataset 1.

Place	Real Period (Hour)	Period[1] (Hour)	Period[2] (Hour)	Period[3] (Hour)	Period[4] (Hour)
Vice-Chancellor Building (10)	168	164	165	165	0
Gym (8)	24	25	24	24	24
RSL (7)	168	171	169	169	169
Holiday Inn (6)	168	163	164	164	0
Post Graduate Center (5)	168	164	165	165	0
WoolWorth Shopping Centre (9)	168	171	170	170	0

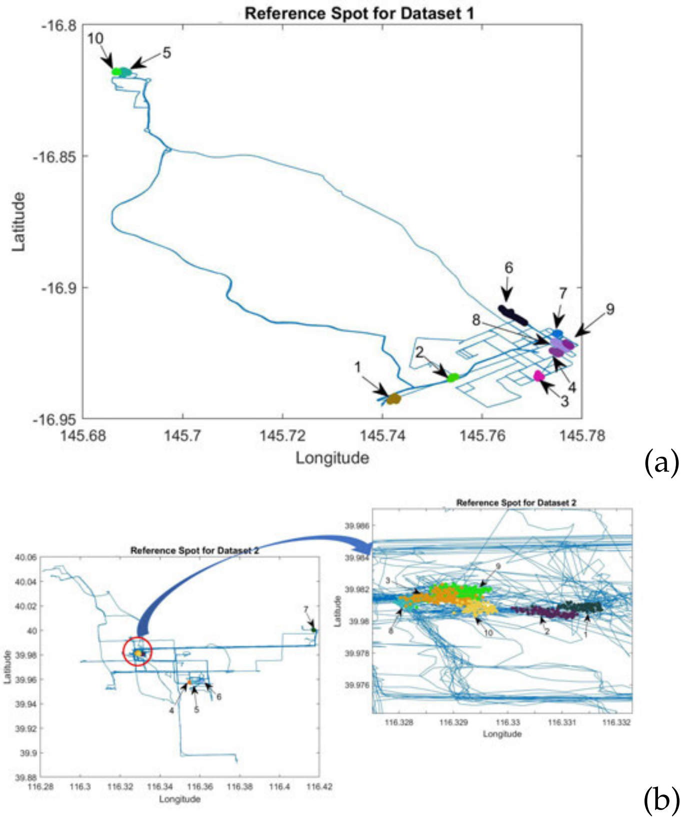
[1] :Lomb-Scargle periodogram
[2] :Fourier transform and autocorrelation (10 seconds)
[3] :Fourier transform and autocorrelation (60 seconds)
[4] :Fourier transform and autocorrelation (120 seconds)

TABLE 5.7: Efficiency comparison of period detection for Dataset 1.

	LSP[1]	FT&A[2]	FT&A[3]	FT&A[4]
Running Time (seconds)	2.52	1422.86	198.48	99.93

[1] :Lomb-Scargle Periodogram
[2] :Fourier transform and autocorrelation (10 seconds)
[3] :Fourier transform and autocorrelation (60 seconds)
[4] :Fourier transform and autocorrelation (120 seconds)

5.3.6 Effectiveness of Periodic Patterns



(a)

(b)

FIGURE 5.19: Visualisations of reference spots for Dataset 1 and Dataset 2.

We use the same method used in Traclus (ST) to find periodic patterns. Table 5.8 shows a number of periodic patterns that Periodica and Traclus (ST) fail to detect for Dataset 1 and Dataset 2. For instance, in Dataset 1, the data owner regularly goes to the toastmaster club (RLS) to practice English in place 7, then goes to a supermarket for shopping in place 9. Both place 7 and place 9 have the same time period (171 hours, almost equivalent to one week), thus, we can get a periodic pattern place 7→0→place 9, where 0 means any point not in the list of reference spots. A periodic pattern place 1→0→place 10 shows that he goes to University (place 10) from home (place 1) for weekly meetings. Note that, there is no period for place 1, but place 1→0→place 10 is still a periodic pattern, because place 10 has an associated regular period. The data owner always goes to place 10 from place 1. Figure 5.19(b) shows reference spots in Dataset 2 where the right figure displays a zoomed area for the red circle in the left figure.

We do not have an associated ground-truth for Dataset 2, but we can infer user's behaviors. For instance, there is a periodic pattern associated with place 1 that is a branch of the Chinese Academy of Sciences (CAS). Our approach finds a regular period with almost 12 hours. It might be this user comes to this place to work in the morning for a day shift and then leaves this building after work, and he/she returns to this place again at night for a night shift. Furthermore, we can find he/she always goes to place 1 (a branch of the Chinese Academy of Sciences, CAS) from place 2 (a conference room), which indicates a regular periodic pattern from a conference room to the branch office. Note that, place 9 is an apartment block, and place 3 is a supermarket. There is a regular period of 71 hours for place 9. That means, he/she goes to place 9 with a 3 days time interval as he/she needs to do something regularly. We can also find that he/she always goes to place 3 (a supermarket) before going to place 9. This could mean that he/she regularly buys things before he/she goes to place 9. The same scenario for place 3→0→ place 8 is found with a 24 hours regular period on place 8. Although there is no annotation on the map regarding this place 8, but we can witness that he/she visits place 8 on a daily basis.

TABLE 5.8: Periodic patterns from Dataset 1 and Dataset 2.

	Reference spot	period (Hours)	Periodic Patterns
Dataset 1	Woolworth (9)	171	7→0→9
Dataset 1	Gym (8)	25	1→0→8
Dataset 1	RSL (7)	171	1→0→7
Dataset 1	Vice-Chancellor Building (10)	164	1→0→10
Dataset 2	CAS (1)	13	2→0→1
Dataset 2	Apartment Block (9)	71	3→0→9
Dataset 2	Building (8)	24	3→0→8

We compare periodic patterns detected by Periodica and Traclus (ST) using 10 seconds as the time interval against those identified by our approach. Although Periodica and Traclus (ST) do not consider background information, we can derive semantic information for some reference spots based on known background information for comparison. Figure 5.20 shows obtained reference spots by Periodica and Traclus (ST) for Dataset 1. Figure 5.21 displays reference spots identified from Periodica and Traclus (ST) for Dataset 2. Table 5.9 shows some periodic patterns from Periodica and Traclus (ST) for Dataset 1 and Dataset 2.

For Dataset 1, only reference spot 2 from Periodica obtains a correct period of 167 hours compared to reference spot 6 in Figure 5.19(a). The periodic pattern 6→0→2 (Periodica) is similar to 1→0→6 (obtained by our method). The remaining reference spots do not match well with the semantic places. For instance reference spot 3, it is too widely defined, and it includes several semantic places, and does not match with one particular semantic place. Also, as with reference spot 4, it is close to semantic place 3, but does not match with it. Interpretation of the periodic patterns of Traclus (ST) for Dataset 1 is not straightforward to interpret them because we cannot find similar reference spots from known semantic places to compare against. As shown in Figure 5.20(b), we see reference spots 3, 4 and 5 as frequent paths, but they do not completely match with known semantic places. Similarly, reference spots 1, 2 and 6 are close to some known semantic places, but they do not coincide with known semantic places.

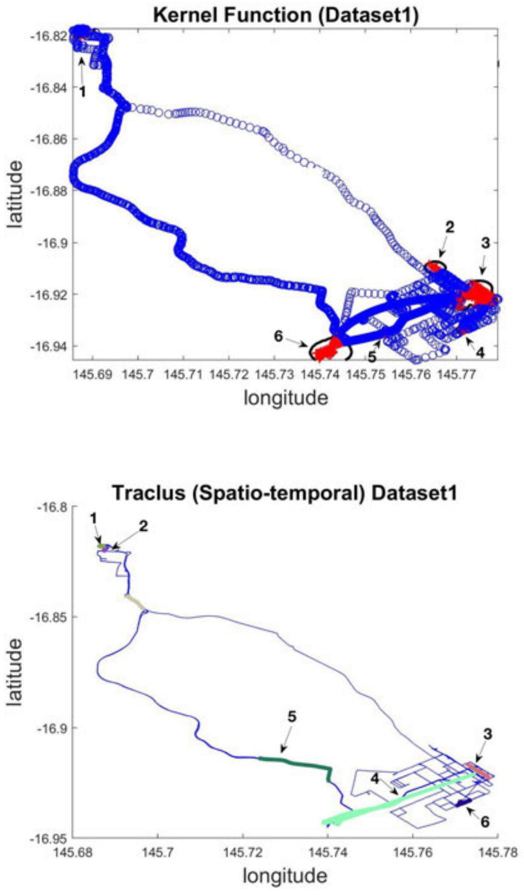
For Dataset 2, Periodica suffers from the same problem as with Dataset 1. In periodic patterns 2→0→1 and 1→0→3, the reference spot 1 is too widely defined and it seems to be equivalent to a combination of reference spots 1, 2, 3, 8, 9 and 10 in our method, and reference spot 3 in Periodica seems to be equivalent to reference spots 4 , 5 and 6 in our method. This means that Periodica misses out local variations and local semantic patterns, thus eventually failing to detect localised periodic patterns. A similar problem occurs with Traclus (ST) that reference spot 1 is too global and widely defined, and it corresponds to reference spots 1, 2, 3, 8, 9, and 10 in our method. Reference spots 3 and 4 cover reference spots 4, 5 and 6 in our method. This means that our approach detects more localised reference spots, and able to generate semantically localised periodic patterns. Whereas Periodica and Traclus (ST) tend to generate large and wide reference spots that do not match with a single semantic place, but rather encompass many semantic places while including false negative semantically meaningless places.

TABLE 5.9: Periodic patterns for Dataset 1 and Dataset 2.

	Reference spot	period (Hours)	Periodic Patterns
Dataset 1 (Periodica)	2	167	6→0→2
Dataset 1 (Traclus (ST))	4	1	4→0→3
Dataset 2 (Periodica)	1	1	2→0→1
Dataset 2 (Periodica)	3	1	1→0→3
Dataset 2 (Traclus (ST))	1	8	2→0→1
Dataset 2 (Traclus (ST))	3	41	1→0→3

5.4 Hierarchical Periodic Patterns

This section focuses on finding hierarchical reference spots for PPM. HDBSCAN is a hierarchical version of traditional DBSCAN which can handle clusters with different varying densities. Thus, we extend DBSCAN to HDBSCAN to find hierarchical reference spots and relevant periodic patterns based on background semantic information. In addition, we still need to consider sequence for each hierarchical reference spot, therefore we employ the same method as in single-level to achieve this: sequential order should increase incrementally and a sudden increase between two points is not permitted.



(a)

(b)

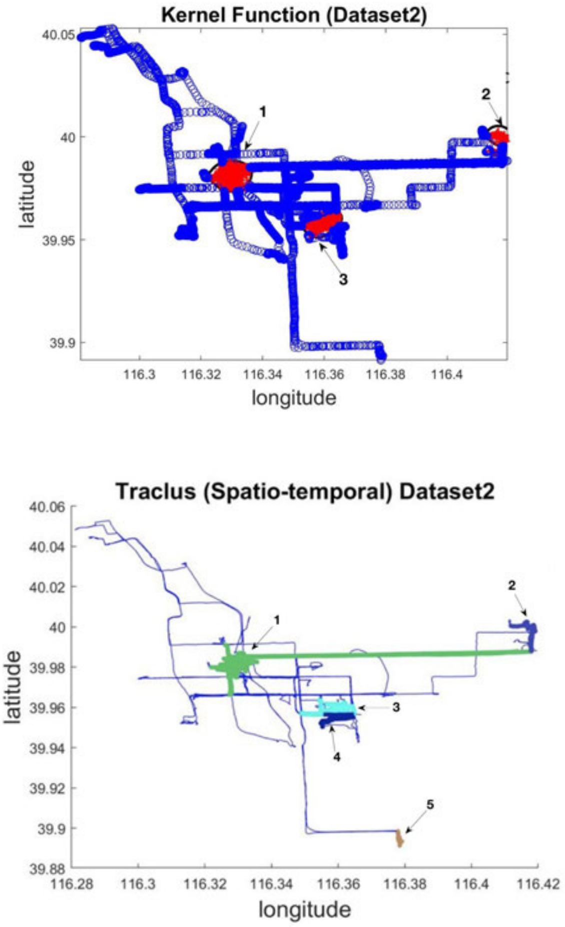
FIGURE 5.20: Visualisation of reference spots from Periodica and Traclus (ST) for Dataset 1: (a) Kernel function; (b) Traclus (ST).

5.4.1 Efficiency

In this section, we compare efficiency on two aspects: one is the whole procedure of both our method and Traclus (ST) since both approaches are hierarchical PPM approaches; the other is period detection methods which used LSP for our approach, and FT&Auto used for Traclus (ST) and Periodica.

Our method vs. Traclus (ST)

Figure 5.22 displays histograms of time efficiency of the whole procedure for single-level and multi-level (hierarchical) between Traclus (ST) and our method with Dataset 1 and Dataset 2. In our method, single-level and multi-level approaches require around 15 seconds and 65 seconds for Dataset 1, and take around 5 seconds and 25 seconds for Dataset 2, respectively. With our approach, the multi-level approach requires slightly more time than the single-level approach, but it is negligible when compared to their difference with Traclus (ST). One important thing to note is that our approach requires



(a)

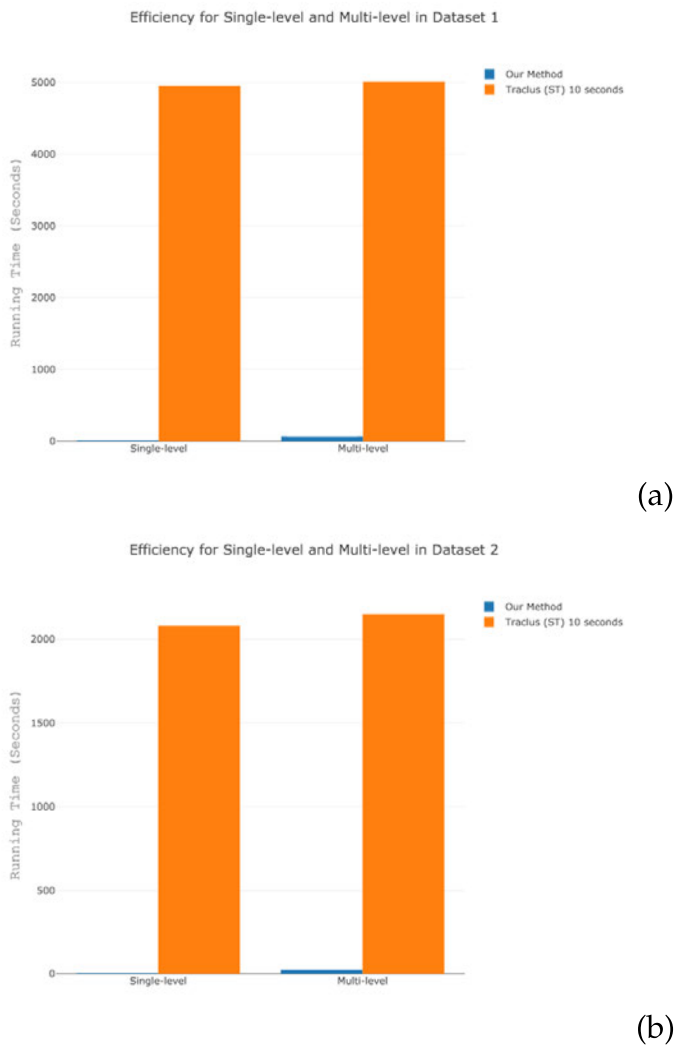
(b)

FIGURE 5.21: Visualisation of reference spots from Periodica and Traclus (ST) for Dataset 2: (a) Kernel function; (b) Traclus (ST).

significantly less time than Traclus (ST) which needs an interpolation process to make irregular trajectories regular. Therefore, our approach is scalable and well suited to GPS-collected spatio-temporal trajectories of a large scale.

Lomb-Scargle Periodogram VS Fourier Transform and Autocorrelation

Figure 5.23 shows histograms of LSP and FT&Auto with 10 seconds and 120 seconds interpolation intervals, respectively. Not surprisingly, LSP not requiring an interpolation process only spends around 3 seconds for period detection on all hierarchical semantic places for Dataset 1 and Dataset 2. Due to the interpolation process, FT&Auto spends 1084.075 and 837.6898 seconds for Dataset 1 and Dataset 2 using a 10 seconds interval, respectively. It still takes more than 45 seconds for Dataset 1 and Dataset 2 even in the use of 120 seconds as a time interval at the expense of reference spot quality (using this large interval will lead to many false negatives). The time efficiency of LSP greatly outperforms that of FT&Auto (with 10 and 120 seconds time intervals) with these two datasets, which confirms that LSP is



(a)

(b)

FIGURE 5.22: Comparison of efficiency for single-level and multi-level between our method and Traclus (ST): (a) Efficiency with Dataset 1; (b) Efficiency with Dataset 2.

more efficient and well suited for GPS-collected irregularly sampled spatio-temporal trajectories.

5.4.2 Effectiveness of Hierarchical Reference Spots

As is the case with single-level, because of the known ground truth in Dataset 1, some effectiveness measures use Dataset 1 to quantitatively compare the performance of our approach against Traclus (ST) and Periodica, while others use Dataset 1 and Dataset 2 to provide more comparative results. As mentioned previously, in single-level, Figure 5.13 displays real semantic places for Dataset 1, and each semantic place is indicated with an associated number and arrow. Background semantic names of places are shown in Table 5.5.

Table 5.10 lists some hierarchical semantic places relevant to the single-level semantic places and also this study. Semantic places 8 & 9 are in

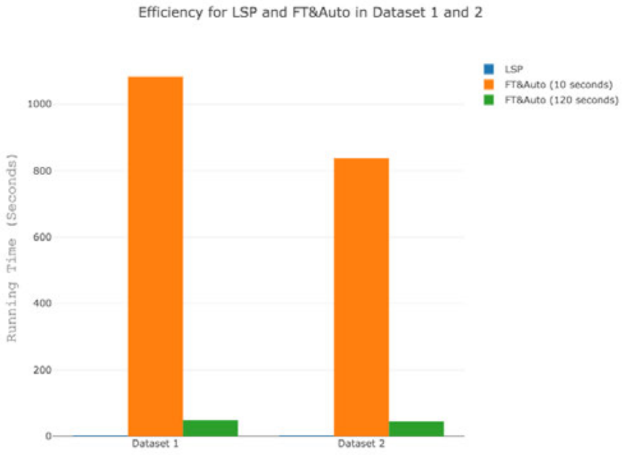


FIGURE 5.23: Comparison of efficiency for LSP and FT&Auto.

'Abbott Street Woolworth Building' where a number of shops are hosted, places 5 & 10 are semantic buildings of 'James Cook University' which is the main tertiary education centre in Cairns, places 8, 9, & 4 are for 'Cairns Rusty Block' which is located in the centre of Cairns city, places 8, 9, 4, & 7 are in Cairns CBD, places 1, & 2 are in 'Earlville Westcourt', places in Cairns CBD and 3 & 6 form 'Inner Cairns City' which becomes 'Outer Cairns City' with places in 'Earlville Westcourt'. All these single-level semantic places are in 'Cairns'.

TABLE 5.10: Names of multi-level semantic places for Dataset 1.

Number	Semantic Places
(8, 9)	Abbott Street Woolworth Building
(5, 10)	James Cook University
(8, 9), (4)	Cairns Rusty Block
(8, 9, 4), (7)	Cairns CBD
(8, 9, 4, 7), (3), (6)	Inner Cairns City
(1, 2)	Earlville Westcourt
(8, 9, 4, 7, 3 , 6), (1, 2)	Outer Cairns City
(8, 9, 4, 7, 3, 6, 1, 2), (5, 10)	Cairns

Figure 5.24 displays single-level reference spots (semantic places) for Dataset 1 and Dataset 2 and their corresponding dendrograms. Figure 5.24(a) and (c) represent visualisations of single-level reference spots for Dataset 1 and Dataset 2, respectively whilst Figure 5.24 (b) and (d) show dendrograms generated by our approach for multi-level semantic places for Dataset 1 and Dataset 2, respectively. These show that our approach not

only identifies single-level semantic places, but various multi-level semantic places as shown in Table 5.10.

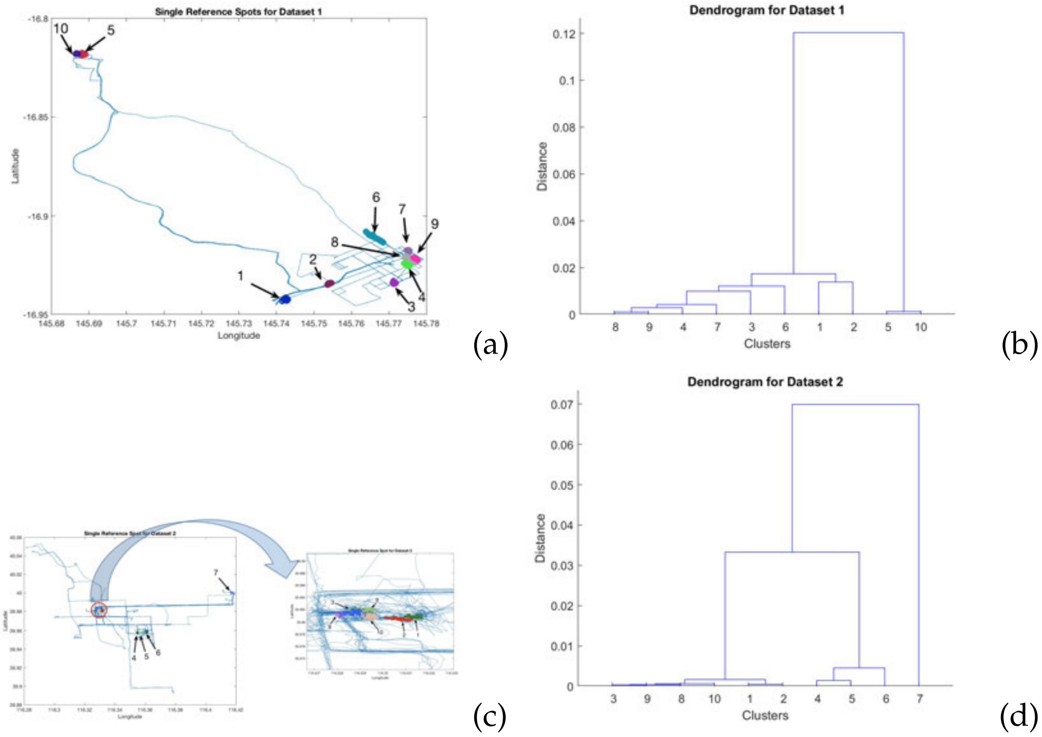


FIGURE 5.24: Dendrograms of our approach for Dataset 1 and Dataset 2.

Figure 5.25 visualise generated single-level reference spots from Periodica for Dataset 1 and Dataset 2. Although Periodica does not produce hierarchical reference spots, we modify it to generate hierarchical reference spots using the single linkage approach to be comparable to our approach.

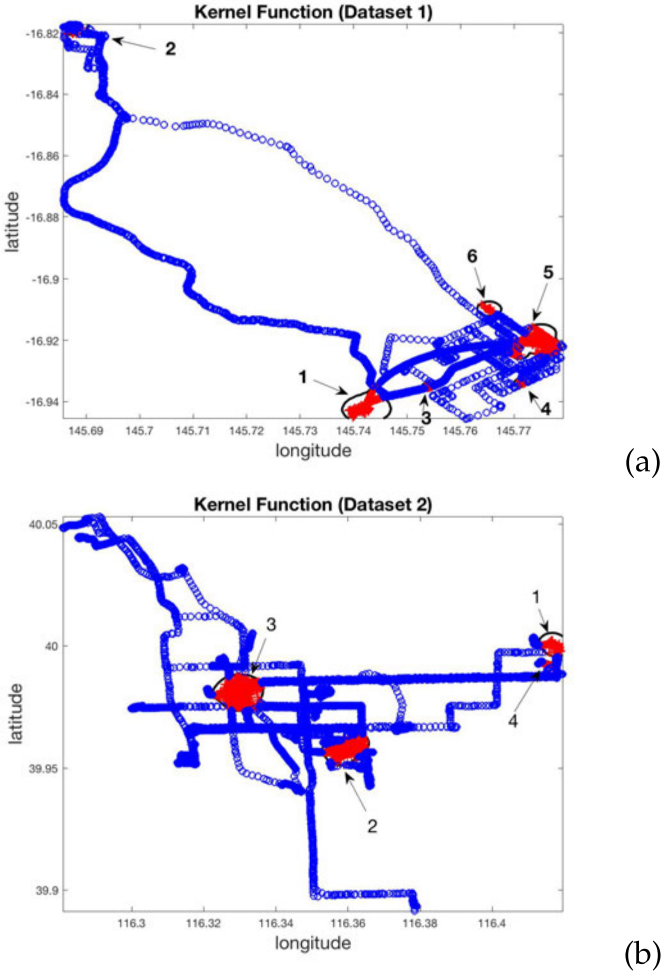
In this section, as with single-level, we compare the effectiveness of our approach against Periodica and Traclus (ST) with regard to: (1) number of reference spots; (2) spatial compactness; (3) temporal compactness; (4) spatio-temporal compactness; (5) semantic accuracy.

Number of Hierarchical Reference Spots

We still use a 10 seconds time interval for interpolation in Periodica and Traclus (ST) to generate a greater number of more fine-tuned single-level reference spots. Table 5.11 shows the number of hierarchical reference spots for Dataset 1 and Dataset 2. It confirms our method can find more hierarchical reference spots than Periodica and Traclus (ST).

Spatial Compactness

In contrast with single-level, the centroid of a hierarchical semantic place is calculated by the mean of centroids of associated semantic places. For example, in Figure 5.24(b), single-level semantic places 8 (gym) and 9



(a)

(b)

FIGURE 5.25: Visulisation of Periodica for Dataset 1 and Dataset 2.

(Woolworth supermarket) are merged to a new hierarchical semantic place (Abbott Street Woolworth Building), and the centroid of the new hierarchical semantic place is calculated by the mean of centroids of semantic places 8 and 9.

Figure 5.26 shows the performance of three methods in spatial compactness. The x-axis shows hierarchical semantic places whilst y-axis shows their corresponding average spatial distance. The larger average distance becomes, the lower the degree of spatial compactness is noted. Not surprisingly, our method shows higher spatial compactness (lower average distance) than Periodica and Traclus (ST) for each hierarchical semantic place. Periodica shows a better performance than Traclus (ST) in spatial compactness. The average spatial compactness of Periodica is 0.0067, Traclus (ST) is 0.0211, whilst our method is 0.0038 which demonstrates the superiority of our method in spatial compactness.

TABLE 5.11: The number of hierarchical reference spots for Dataset 1 and Dataset 2.

Approach	Dataset 1	Dataset 2
Periodica (10 seconds)	5	3
Traclus (ST) (10 seconds)	6	4
Our Method	9	9

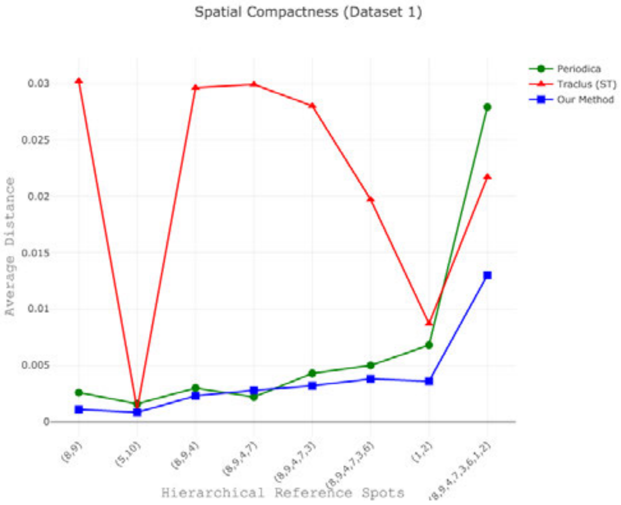
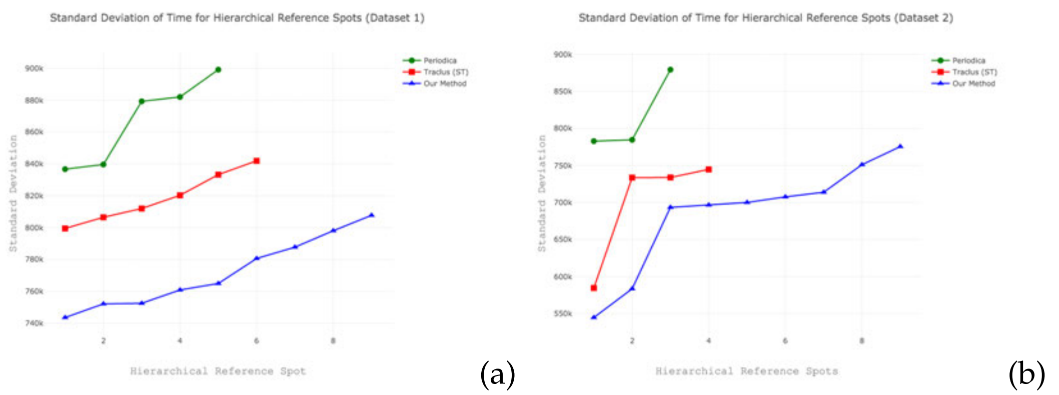


FIGURE 5.26: Measure of spatial compactness with the three methods for Dataset 1 (average spatial distance: 0.0067 for Periodica; 0.0211 for Traclus (ST); 0.0038 for our method).

Temporal Compactness

Hierarchical semantic places should also be spatially close and also temporally aggregated. We use the standard deviation to measure temporal compactness for Dataset 1 and Dataset 2. Figure 5.27 shows measures of temporal compactness with Periodica, Traclus (ST) and our method. Note that, the standard deviation values are still sorted by an ascending order for better comparison and visualisation. The x-axis represents the number of hierarchical semantic places in each method whilst the y-axis shows a corresponding standard deviation value for each hierarchical semantic place. The higher the standard deviation value becomes, the lower the degree of temporal compactness is noted. Although both Traclus (ST) and our method consider the sequence of trajectory, our method achieves a higher temporal compactness value (lower standard deviation) for most hierarchical semantic places. Periodica ignores the sequence of trajectory, not surprisingly, and it shows the lowest temporal compactness. Figure 5.27(a) shows a temporal compactness graph for Dataset 1, where the average of standard deviation is 867370 for Periodica, 818910 for Traclus (ST) and 772030 for our method. In Dataset 2, the average of standard



(a) (b)

FIGURE 5.27: Measure of temporal compactness with the three methods for Dataset 1 and Dataset 2 (average temporal distance: 867370 for Dataset 1 and 815600 for Dataset2 for Periodica; 818910 for Dataset 1 and 699160 for Dataset 2 for Traclus (ST); 772030 for Dataset 1 and 685128 for Dataset 2 for our approach).

deviation is 815600 for Periodica, 699160 for Traclus (ST) and 685128 for our method as shown in Figure 5.27(b). Clearly our method has a better performance than Periodica and Traclus (ST) in temporal compactness.

Spatio-temporal Compactness

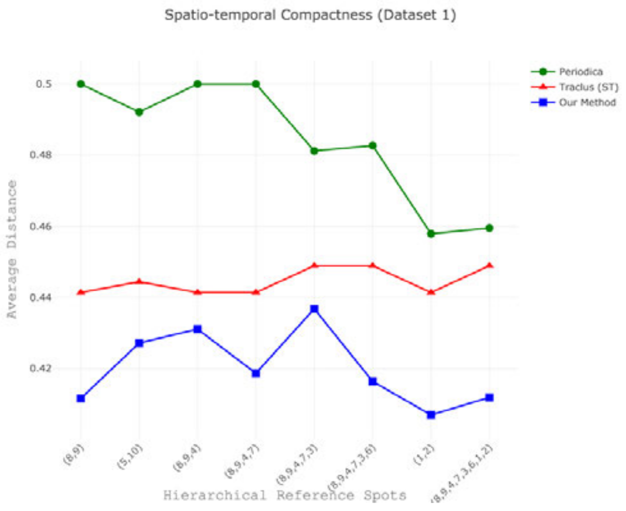


FIGURE 5.28: Measure of spatio-temporal compactness with the three methods for Dataset 1 (average spatio-temporal distance: 0.4842 for Periodica; 0.4446 for Traclus (ST); and 0.4201 for our approach).

Similar to sing-level, due to different scales in spatial compactness and temporal compactness, we normalise them into the same scale [0,1] using min-max normalisation to handle the different scale effect. Space and time still have the same weight (0.5).

Figure 5.28 displays spatio-temporal compactness values of three methods where the x-axis shows semantic places whilst the y-axis represents a sum of normalised standard deviation value (inverse TC) and normalised average distance (inverse SC). The higher sum average distance value becomes, the lower the degree of STC value is noted in the y-axis. Our method obtains the lowest distance value for each hierarchical semantic place when compared to Periodica and Traclus (ST). Note that, although Traclus (ST) obtains lower spatial compactness than Periodica, Traclus (ST) gains higher spatio-temporal compactness than Periodica, which indicates Traclus (ST) achieves a better performance than Periodica in spatio-temporal compactness. The average spatio-temporal distance values for Dataset 1 are 0.4842 for Periodica, 0.4446 for Traclus (ST) and 0.4201 for our method. Our method is superior to Periodica and Traclus (ST) in spatio-temporal compactness.

Semantic Accuracy

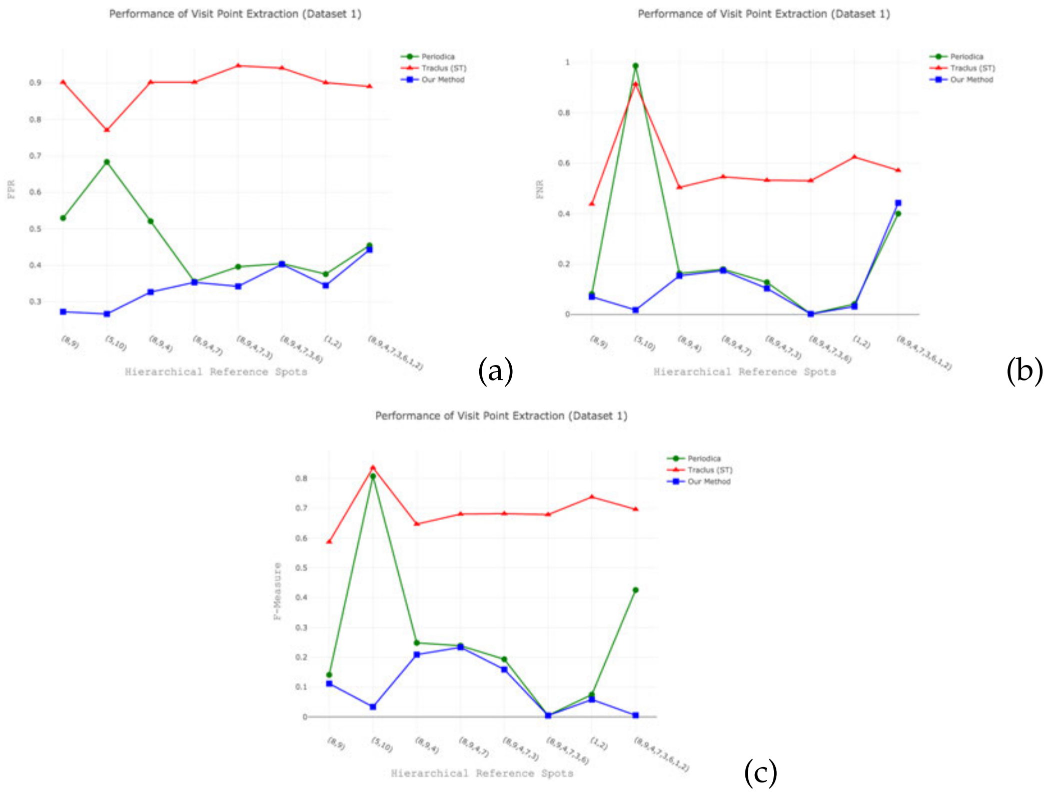


FIGURE 5.29: Performance comparison of the three methods for place extraction: (a) FPR graph; (b) FNR graph; (c) F-measure graph.

In this section, we still measure semantic accuracy by False Positive Ratio (FPR), False Negative Ratio (FNR), and F- measure for hierarchical reference spots. Figure 5.29 (a), (b) and (c) display FPR, FNR and F-measure of three algorithms for Dataset 1, respectively. Not surprisingly, our method obtains a better performance than Periodica and Traclus (ST) in FPR, FNR and F-measure for each hierarchical semantic place.

5.4.3 Effectiveness of Periods for Hierarchical Semantic Places

TABLE 5.12: Obtained periods for some semantic places for Dataset 1.

Place	Obtained Period (Hour)
Vice-Chancellor Building (10)	164
Gym (8)	25
RSL (7)	171
Holiday Inn (6)	163
Post Graduate Centre (5)	164
WoolWorth (9)	171

TABLE 5.13: Obtained periods for some semantic places for Dataset 2.

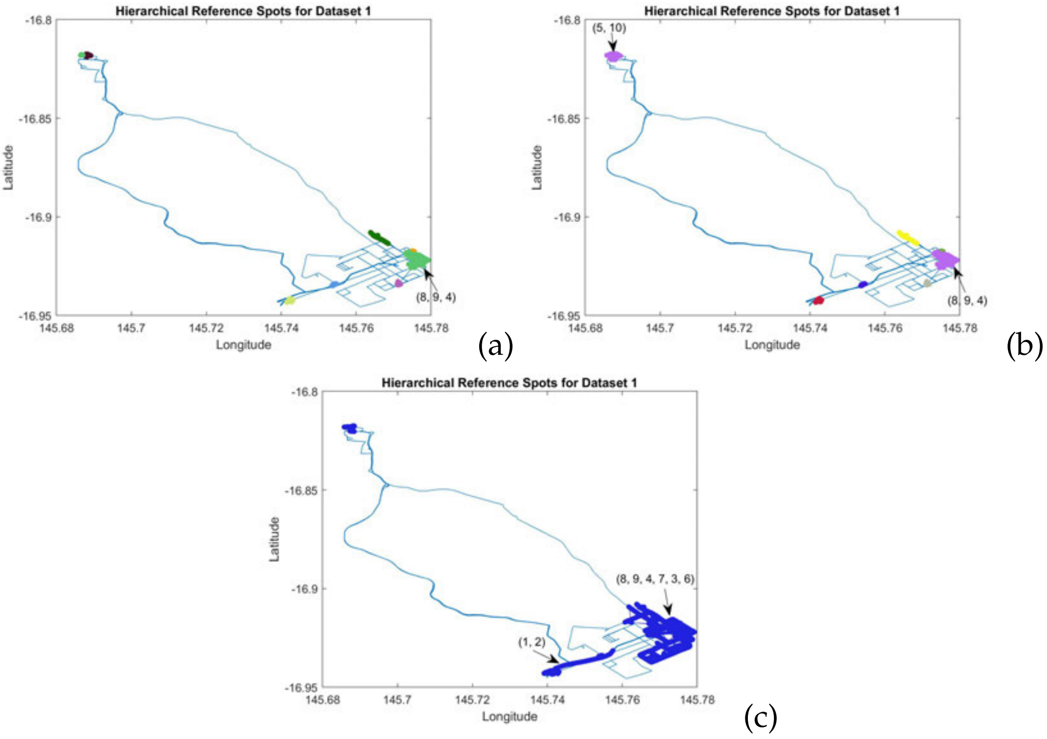
Place	Obtained Period (Hour)
A branch of Chinese Academy of Sciences (1)	13
An apartment block (9)	71
A building (8)	24
A supermarket (3)	23

Table 5.12 and Table 5.13 show some single-level semantic places which are matched into different semantic places with non-zero periods in Dataset 1 and Dataset 2, respectively. Obviously, not all semantic places have non-zero periods, thus, there are no periodic patterns for some semantic places which have zero periods. For instance, there are no non-zero periods for semantic place 4 in Dataset 1 and semantic place 5 in Dataset 2, thus there are no periodic patterns for both semantic places. As we mentioned before, if we consider the hierarchy of space, there might be some hierarchical semantic places with non-zero periods. After HDBSCAN, we can obtain hierarchical semantic places, and if these hierarchical reference spots are in some larger semantic places, such as suburbs and cities, we can match these reference spots to those larger semantic places.

Table 5.14 shows some hierarchical semantic places and corresponding non-zero periods. For instance, in Dataset 1, single-level semantic place 4 has no period. In Figure 5.30(a), semantic place 4 (Rusty Sunday Market) is

TABLE 5.14: Periods for hierarchical semantic places in Dataset 1 and Dataset 2.

Place	Period (Hour)
Fig. 5.30 (a) : (8, 9) Dataset 1	23
Fig. 5.30 (b) : (8, 9, 4) Dataset 1	23
Fig. 5.30 (b) : (5, 10) Dataset 1	163
Fig. 5.30 (c) : (8, 9, 4, 7, 3, 6) Dataset 1	21
Fig. 5.30 (c) : (1, 2) Dataset 1	25
Fig. 5.31 (a) : (3, 8, 9, 10, 1, 2) Dataset 2	25
Fig. 5.31 (b) : (4, 5) Dataset 2	175
Fig. 5.31 (c) : (4, 5, 6) Dataset 2	26



(a)

(b)

(c)

FIGURE 5.30: Visualisations of hierarchical semantic places using our approach for Dataset 1.

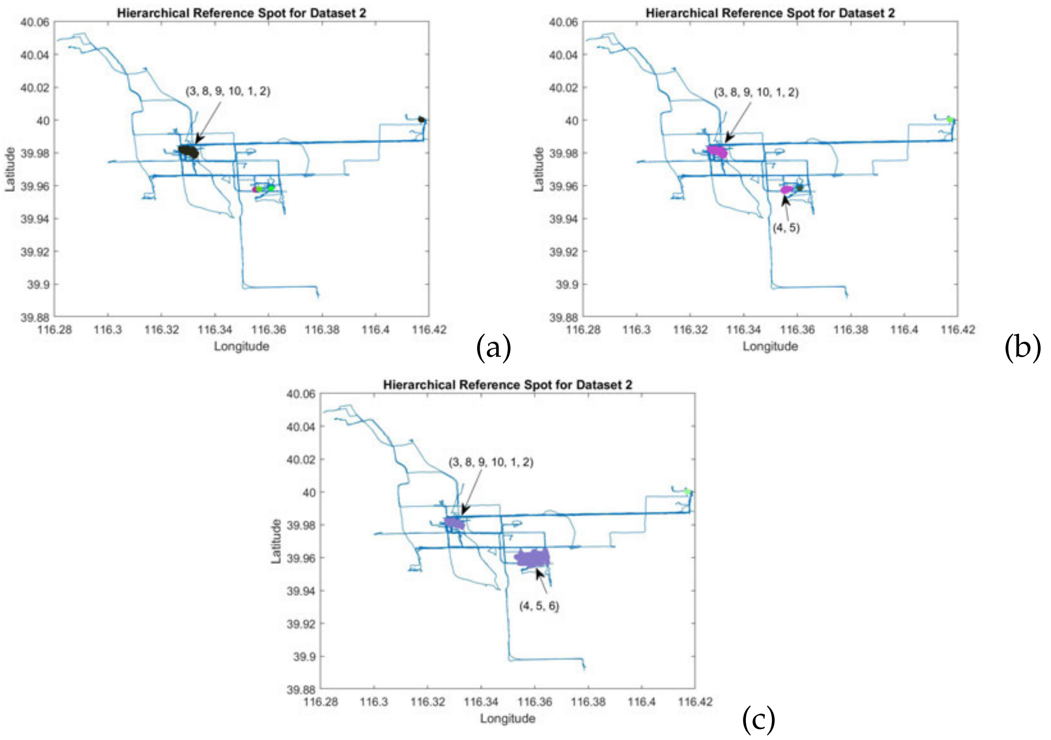


FIGURE 5.31: Visualisations for hierarchical semantic places using our approach for Dataset 2.

combined with semantic places (8, 9) (Abbott Stree Woolworth Buidling) in the dendrogram forming a hierarchical semantic place, ‘Cairns Rusty Block’ (8, 9, 4), and there is a period of 23 hours. This indicates there is a periodic pattern to ‘Cairns Rusty Block’ with a period of 23 hours. Note that single-level semantic place 5 and semantic place 10 shown in Figure 5.30(b), have the same period of 164 hours, thus, the hierarchical semantic place (5, 10) forming a hierarchical semantic place called ‘James Cook University’, obtains a period of 163 hours which is very similar to the period of single-level semantic places 5 and 10. This indicates that the data owner lives in Cairns city, and comes to the university once a week around every 163 hours. In Figure 5.30(c), single-level semantic places 1 and 2 can be combined as a new hierarchical semantic place, ‘Earlville Westcourt’. Although there is no period for each of these single-level semantic places, a period of 25 hours is obtained for ‘ Earlville Westcourt’. This implies that the data owner periodically comes to ‘Earlville Westcourt’ where the data owner’s home is located, regularly with around 25 hours period. Note that it is not possible to detect these periodic patterns with hierarchical semantic places using traditional approaches, but our proposed hierarchical approach is able to do so.

In Dataset 2, a set of single-level semantic places (3, 8, 9, 10, 1, 2) forms ‘Zhongguancun Campus of the University of Chinese Academy of Sciences’, semantic places (4, 5) form ‘Academic Buildings and Restaurants’, whilst semantic places (4, 5, 6) form ‘Beijing Normal University’. In Figure 5.31(a), single-level semantic places 2 and 10 with zero periods can be combined with semantic places 3, 8, 9, and 1 as a new semantic place (3, 8, 9, 10, 1, 2), which is with a period of 25 hours. Figure 5.31(b) and (c) show single-level semantic places 4, 5 and 6 can be combined as two new semantic places (4, 5) or (4, 5, 6). Although there are no periods for single-level semantic places 4, 5 and 6, two periods of 175 hours and 26 hours can be obtained after merging them.

5.4.4 Hierarchical Periodic Patterns

TABLE 5.15: Hierarchical periodic patterns from Dataset 1 and Dataset 2.

	Semantic Place	Period (Hours)	Periodic Patterns
Dataset 1	(5, 10)	163	(1)→0→(5, 10)
Dataset 1	(8, 9, 4, 7, 3, 6)	21	(1, 2)→0→(8, 9, 4, 7, 3, 6)
Dataset 2	(4, 5)	175	(6)→0→(4, 5)
Dataset 2	(3, 8, 9, 10, 1, 2)	25	(4, 5, 6)→0→(3, 8, 9, 10, 1, 2)
Dataset 2	(4, 5, 6)	26	(3, 8, 9, 10, 1, 2)→0→(4, 5, 6)

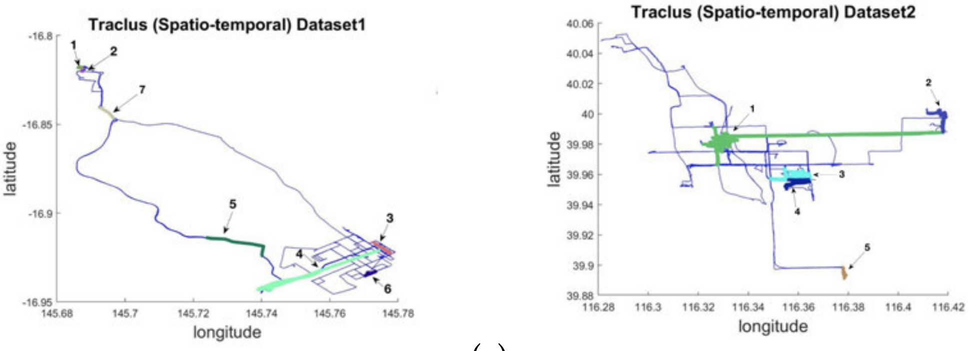
We use the same approach in Chapter 4 to find periodic patterns. Table 5.15 shows some interesting periodic patterns based on hierarchical semantic places for Dataset 1 and Dataset 2. In Dataset 1, we know the ground truth, thus we can interpret the periodic patterns with relative ease. Semantic places 5 and 10 can be combined as the whole university campus with a period of 163 hours. A periodic pattern (1)→0→(5, 10) means that the data owner has a weekly meeting, thus he goes to university once every week from his place (semantic place 1). 0 means the data owner is not in any semantic place. Note that, there are no periods for single-level semantic places 1, 2, 3 and 4. However, if we consider hierarchical semantic places, semantic places (1, 2) and semantic places (8, 9, 4, 7, 3, 6) can be merged to form two new hierarchical semantic places, Earlville Westcourt and Inner Cairns City, respectively. There is a periodic pattern (1, 2)→0→(8, 9, 4, 7, 3, 6) with a period of 25 hours between them. (1, 2) is a part of Earlville Westcourt, which is a suburban area of Cairns. (8, 9, 4, 7, 3, 6) is covered by Inner Cairns City. Thus, this periodic pattern actually reveals a repeating movement between (Earlville Westcourt) and (Inner Cairns City) with a period of 21 hours.

In reality, considering hierarchical semantic places are of use and importance in PPM, because (1) a single-level semantic place might have a zero period, we cannot find a periodic pattern for the single-level semantic place. However there could exist a non-zero period for higher level hierarchical semantic places such as a periodic pattern (1, 2)→0→(8, 9, 4, 7, 3, 6). This kind of periodic pattern is of great use but cannot be found with traditional approaches; (2) we can use these hierarchical periodic patterns for better decision-making. For example, urban planners are able to make an overall plan for allocating limited resources to different areas according to people's behaviors. Obviously, periodic patterns for hierarchical semantic places can help urban planners allocate limited resources fairly across the region based on people's periodic movements. In Dataset 2, semantic places 4 and 5 can be combined as a new hierarchical semantic place with a period of 175 hours. Note that, in Table 5.13, there are no periods for single-level semantic places 4 and 5. Semantic places 4 and 5 are annotated as buildings, semantic place 6 is a teaching building, thus, we infer that a periodic pattern (6)→0→(4, 5) shows the object went to semantic place 4 or 5 for a certain aim (e.g. visiting friends) after a class in semantic place 6. Another periodic pattern (4, 5, 6)→0→(3, 8, 9, 10, 1, 2) shows the object regularly goes to (3, 8, 9, 10, 1, 2) with a period of 25 hours, (3, 8, 9, 10, 1, 2) is the University of Chinese Academy of Sciences and (4, 5, 6) is Beijing Normal University, thus we infer this person might be a teacher or a student, because he/she always moves from (4, 5, 6) (Beijing Normal University) to another place (3, 8, 9, 10, 1, 2) (Zhongguancun Campus of the University of Chinese Academy of Sciences) and back again.

5.4.5 Comparing Traclus (ST) for Hierarchical Semantic spots in PPM

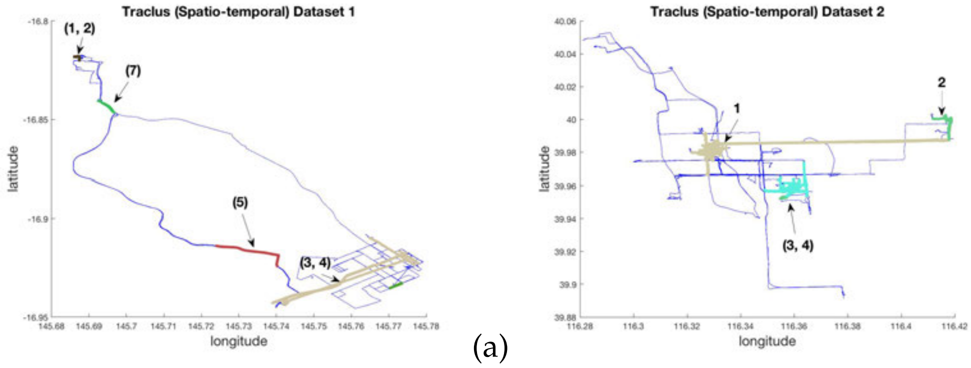
Even though Traclus (ST) can effectively find hierarchical reference spots for PPM, this method fails to consider background semantic information. In this section, we compare hierarchical reference spots from our method and hierarchical reference spots from Traclus (ST) for PPM with Dataset 1 and Dataset 2. For a fair comparative study, we extract background semantic information based on our approach for reference spots from Traclus (ST). Figure 5.32 shows obtained single-level reference spots from Traclus (ST). For Dataset 1, reference spots 3, 4 and 5 match with some road segments instead of any meaningful semantic places. Reference spots 1, 2 and 6 seem to only partially match with some semantic places, but not entirely match with them. Therefore, hierarchical reference spots from Traclus (ST) do not correctly represent semantic places, and thus fail to generate semantic periodic patterns. Traclus (ST) for Dataset 2, reference spots 1 and 2 match with some well-defined roads, and reference spots 3 and 4 seem to match with Beijing Normal University. Figure 5.33 displays some hierarchical reference spots from Traclus (ST) for Dataset 1 and Dataset 2. Table 5.16 also shows some periodic patterns which are relevant to these hierarchical reference spots. In Dataset 1, a periodic pattern 7→0→(1, 2) is not useful, because reference spot 7 is not a semantic place, but it is just a section of road. There is no special meaning for this reference spot. In addition, although a hierarchical reference spot (1, 2) seems to match with the university, there is a period of one hour for this pattern which is incorrect. This is due to the fact that Traclus (ST) cannot effectively reveal semantic places, but rather focuses on repeatedly visited paths. A periodic pattern

5→0→(3, 4) suffers from the same problem, that is, reference spot 5 is a section of road. For a periodic pattern (3, 4)→0→(1) in Dataset 2, our method can find a correct semantic place (Zhongguancun Campus of the University of Chinese Academy of Sciences), whereas, reference spot 1 from Traclus (ST) cannot.



(a) (b)

FIGURE 5.32: Visualisations for single-level reference spots from Traclus (ST) in Dataset 1 and Dataset 2.



(a) (b)

FIGURE 5.33: Visualisations for hierarchical reference spots from Traclus (ST) in Dataset 1 and Dataset 2.

TABLE 5.16: Obtained periods for some reference spots using Traclus (ST) for Dataset 1 and Dataset 2.

	Reference Spot	Period (Hour)	Periodic Pattern
Dataset 1	(1, 2)	1	7→0→(1, 2)
Dataset 1	(3, 4)	5	5→0→(3, 4)
Dataset 2	(1)	2	(3, 4)→0→(1)
Dataset 2	(3, 4)	8	(1)→0→(3, 4)

5.5 Summary

Mining periodic patterns from spatio-temporal trajectories is of great importance since it reveals interesting and regular periodic behaviours. There is a growing interest in efficient and effective PPM from spatio-temporal trajectories due to the wide availability of automatic location collecting devices. In this chapter, we identify the five special characteristics (sequence, spatio-temporality, semantics, hierarchical nature and irregularity) of GPS-collected spatio-temporal trajectories for PPM.

We propose a new node-based hierarchical semantic PPM approach especially designed for spatio-temporal trajectories to successfully meet the five requirements. Our approach first transforms raw spatio-temporal trajectories into semantically enhanced stop episode annotated trajectories by considering aspatial semantic background information from OpenStreeMap. Our approach applies Lomb-Scargle periodogram to irregular trajectories in order to efficiently handle irregular trajectories. Our approach employs HDBSCAN to find hierarchical reference spots. We conducted various experiments with two real datasets to demonstrate the efficiency and effectiveness of our proposed approach. We used Geolife dataset for exploratory mining to identify semantically interesting periodic patterns whilst we collected our own dataset with ground-truth annotations (this is due to the unavailability of ground-truth datasets) and used it for confirmatory mining to evaluate the effectiveness of our approach with known periodic patterns. In multi-level, our proposed approach is able to detect interesting hierarchical periodic patterns that cannot be detected by traditional approaches. These hierarchical periodic patterns could be potentially used for further cause-effect analysis, decision-making, hypothesis generation and intelligent planning.

Chapter 6

Case study : Multi-level Medical Periodic Patterns from Human Movement Behaviors

Human movement behaviors could reveal many interesting medical patterns. Due to the advances in location-aware devices, a large volume of human movement behaviors has been captured in the form of spatio-temporal trajectories. These spatio-temporal trajectories could be used to identify those people who periodically visit medical centres for treatments (patients), working (health professionals) or other purposes. In this chapter, we introduce a medical PPM framework that utilises spatio-temporal hierarchical PPM approaches to find single-level and multi-level medical periodic patterns. We utilise a widely used Geolife dataset to test the feasibility and applicability of our framework. Section 6.1 introduces the background of this case study. In Section 6.2, we review previous studies about medical periodic pattern. Section 6.3 illustrates the framework and propose approaches for medical PPM from aspatial-temporal trajectory. Experimental results are demonstrated in Section 6.4. Finally, the last section sets out conclusion for this case study.

6.1 Introduction

A large number of spatio-temporal trajectories delivers a new opportunity to analyse the behavior of human movements, and it is a solid candidate to distinguish those people who regularly visit medical centres for treatments (patients) and for working (health professionals) from those who not. Therefore, it could be used to identify a set of patients or health professional from massive trajectories in order to develop micro marketing or to further derive periodic patterns from these identified trajectories. For instance, if a person who periodically visits a medical centre at 10am every Saturday for a month could be seen as a patient whilst if a person who regularly comes to the medical centre at 9am everyday could be a health professional working in the place. Once these patients and health professionals are identified, then they can be further mined to reveal periodic patterns.

Spatio-temporal PPM could identify who are health and medical centre related personnel such as patients or health professionals using their corresponding trajectories, and to find multi-level medical periodic patterns that reveal valuable medical behaviours. There are two main groups in the domain of medical PPM. One is general PPM, and the other is spatio-temporal PPM. The former includes PPM in event/sequence (Cao et al., 2004; Huang and Chang, 2004), time series (Berlingerio et al., 2007; Froelich and Wakulicz-Deja, 2009; Jiawei Han, 1999; Ilayaraja and Meyyappan, 2013; Sheng et al., 2006; Yang et al., 2000; Zhang et al., 2007; L. Zhu et al., 2012) and social networks data (Halder et al., 2017; Parthasarathy et al., 2006) whilst the latter involves spatio-temporal trajectories (Cao et al., 2007; Jindal et al., 2013; Li et al., 2010a; Li et al., 2012; Li and Han, 2014). Traditional studies in PPM in medical contexts fall in the first category, and they mine periodic patterns from health time series datasets (Berlingerio et al., 2007; Froelich and Wakulicz-Deja, 2009; Ilayaraja and Meyyappan, 2013), thus they fail to mine medical patterns from spatio-temporal trajectories. In Chapter 5, we showed that our approaches can handle irregularly sampled and noisy GPS-collected trajectories, and to mine multi-level hierarchical periodic patterns. In this chapter, we propose a medical PPM framework that utilises cutting-edge spatio-temporal PPM approaches to identify a set of trajectories that exhibits periodic visits to medical centres, and also find hierarchical medical periodic patterns.

6.2 Literature Review

There are two major groups in the domain of medical PPM: 1) general PPM, and 2) spatio-temporal PPM. Past studies in PPM in medical contexts are based on time series datasets (Berlingerio et al., 2007; Froelich and Wakulicz-Deja, 2009; Ilayaraja and Meyyappan, 2013). Ilayaraja and Meyyappan (2013) proposed a method to find frequent occurring diseases in specific geographical area at a given time period using Apriori-based technique. Berlingerio et al. (2007) applied time annotated sequences to discover associative frequent patterns for describing trends of different biochemical variables along the time dimension. Froelich and Wakulicz-Deja (2009) applied Fuzzy Cognitive Maps (FCMs) to extract medical concepts from temporal diabetes data for mining periodic frequent patterns. One common drawback with these approaches is that they deal with time series data considering the temporal dimension, but fail to consider the spatial dimension that indicates 'where' periodic patterns occur. Also, these traditional approaches are limited to single-level patterns ignoring the inherent hierarchical nature of patterns. In data-rich medical settings, it is important to effectively find which trajectory (user movement) is stopping at and visiting to medical centres, and what are their multi-level medical periodic patterns. For instance, a user 'A' is visiting to a clinic at 10am for an hour every morning for a period of one month could indicate the user 'A' has a regular medical treatment everyday for a month to cure a certain disease. Past studies with time series data cannot find this kind of spatiality associated periodic pattern.

Table 6.1 compares traditional PPM approaches for medical patterns from spatio-temporal trajectories. Traclus (ST) and semantics-based approach are able to detect hierarchical patterns and both equally and simultaneously consider spatiality and temporality, thus these two approaches will be used for this study.

TABLE 6.1: Comparison of traditional PPM approaches.

	Fixed period	Reference spot	Traclus (ST)	Semantics
Spatio-temporal	✓	×	✓	✓
Irregularity	×	×	×	✓
Hierarchical	×	×	✓	✓
Medical semantics	×	×	×	✓
Sequence	×	×	✓	✓

6.3 Medical PPM Framework

Figure 6.1 depicts an overall framework of our multi-level medical PPM from spatio-temporal trajectories. The framework first takes a set of GPS-collected spatio-temporal trajectories, and utilises two spatio-temporal PPM approaches: the path-based approach in Chapter 4 and the node-based approach in Chapter 5 to identify a subset of trajectories that periodically visit medical centres, and their corresponding multi-level medical periodic patterns. In this chapter, a modified version of the path-based approach for medical PPM is referred to as a spatio-temporal dominant approach whilst that of the node-based approach is referred to as a semantics-dominant approach in this chapter.

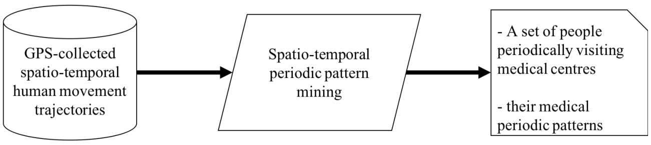

GPS-collected spatio-temporal human movement trajectories
Spatio-temporal periodic pattern mining
- A set of people periodically visiting medical centres
- their medical periodic patterns

FIGURE 6.1: Overall framework of medical PPM from spatio-temporal trajectories.

Algorithm 4 shows a modified spatio-temporal dominant approach for medical PPM whilst Algorithm 5 displays a modified semantics-dominant approach. In semantics-dominant approach, we first extract medical centres from OpenStreetMap and then apply our approach in Chapter 5 to find periodic patterns for medical centres. Different from initial spatio-temporal dominant approach and initial semantics-dominant approach, we only extract medical centres from OpenStreetMap for this study which means our method can extract different semantic places according to different applications. Lines 8-11 in Algorithm 4 and Algorithm 5 extract reference spots that contain medical centres for our study.

Algorithm 4 Spatio-temporal Dominant Approach

```
INPUT: A spatio-temporal trajectory $Traj$, ($\langle x_1,y_1,t_1\rangle$, $\langle x_2,y_2,t_2\rangle,\ldots,\langle x_m,y_m,t_m\rangle,\ldots,\langle x_n,y_n,t_n\rangle$, and a set $M = \{m_1,\ldots,m_k\}$ of medical centres;
OUTPUT: A set of medical periodic patterns;
/* make spatio-temporal trajectory with regular time interval */
Employ Linear interpolation to get the trajectory with a regular time interval, $t_i - t_{i-1} = t_j - t_{j-1}$ for $i \neq j \in \{1,\ldots,n\}$;
/* Find reference spots */
Extend Traclus to additional three implicit trajectory properties $\langle$Direction, Speed, Time$\rangle$ to find reference spots $R = \{r_1, r_2, \ldots, r_j\}$;
/* Extract medical centres from background maps */
Build $M$ from background semantic maps;
/* Detect periods */
for each reference spot $r_i \in R$ do
    if $r_i$ contains any $m_j \in M$ then
        Detect periods for each reference spot $r_i$, and store the periods in $T_i$;
    end if
end for
/* Find periodic patterns */
for each $t \in T_i$ do
    $p_t = \{p_i \mid t \in T_i\}$;
    Construct a symbolised sequence $Q$ using $p_t$;
    Mining periodic patterns from $Q$;
end for
```

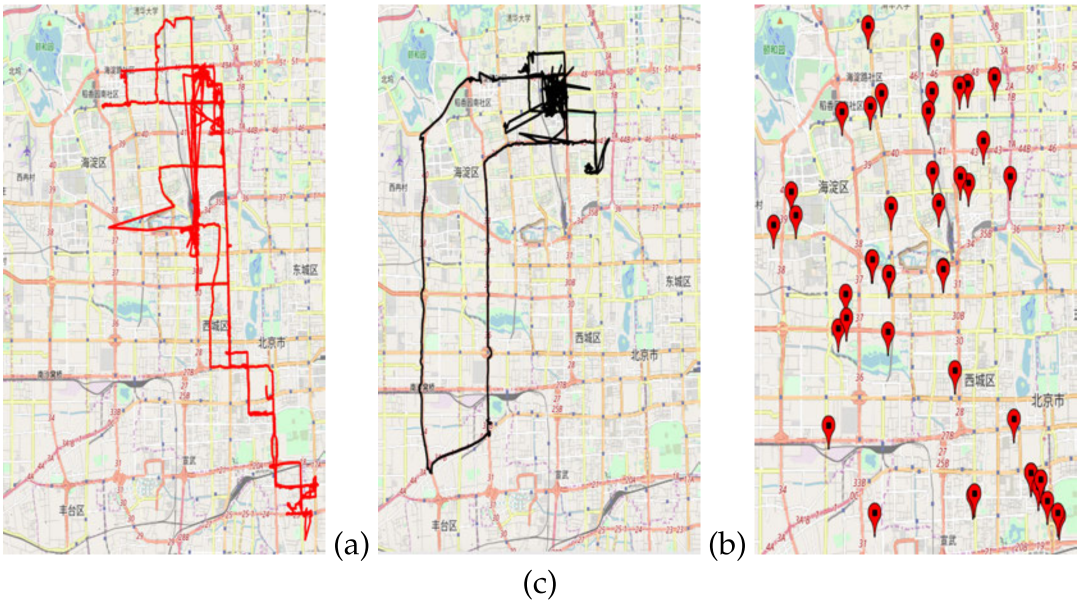
6.4 Experimental Results

6.4.1 Dataset

We use a real GPS dataset from Geolife and Figure 6.2(a) displays one red trajectory, which is recorded from 26/9/2008 to 10/10/2008, has periodic visits to medical centres (referred to as a positive trajectory in this thesis) whilst the other, in black recorded from 25/10/2008 to 10/11/2008, does not have periodic visits to medical centres in Figure 6.2(b)(referred to as a negative trajectory). A set of medical centres in the study region is shown in Figure 6.2(c).

Algorithm 5 Semantics Dominant Approach

```
INPUT: A spatio-temporal trajectory $Traj$, and a set $M$ of medical centres;
OUTPUT: A set of periodic patterns with associated places;
/* Find stopping places using HMM */
Find stop episodes $S = \{s_1, s_2, ..., s_n\}$;
/* Map matching those stopping episodes to real places */
for each $s_i \in S$ do
    Match each stop episode in $S$ to places $P = \{p_1, p_2, ..., p_n\}$;
end for
/* Detect periods for each stopping place */
for each place $s_i \in P$ do
    if $s_i$ contains any $m_j \in M$ then
        Detect periods for $p_i$ that matches with $s_i$, and store the periods in $T_i$;
    end if
end for
/* Mine periodic patterns */
for each $t \in T_i$ do
    $p_t = \{p_i \mid t \in T_i\}$;
    Construct a symbolised sequence $Q$ using $p_t$;
    Mining periodic patterns from $Q$;
end for
```



(a) (b) (c)

FIGURE 6.2: Visualisations of two user trajectories and medical centres: (a) A positve trajectory; (b) A negative trajectory (c) Locations of medical centres.

6.4.2 Efficiency

In this section, we compare the efficiency of the whole procedure for both methods based on the real dataset. Figure 6.3 displays the efficiency of the whole procedure for both methods. The left shows the running time of the semantic dominant approach, the right presents the running time of the spatio-temporal dominant approach. Obviously, the spatio-temporal dominant approach spends 1817.557429 seconds whilst the semantic dominant approach takes 5.060988 seconds for the whole procedure. Obviously, the latter is much more efficient than the former. The main reason is that the spatio-temporal dominant approach needs interpolation to make irregular raw trajectories regular for subsequent period detection. Note that, the semantic dominant approach does not need interpolation since it is able to handle irregular trajectories for period detection.

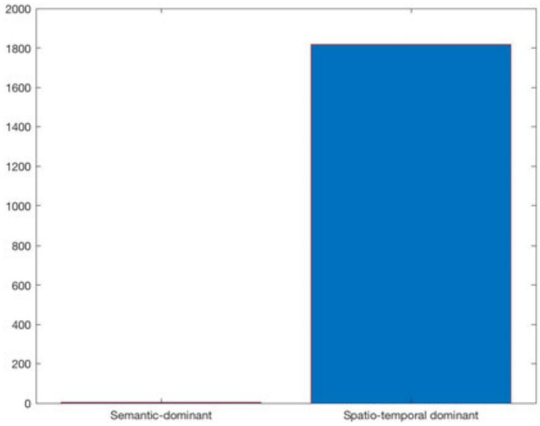


FIGURE 6.3: Efficiency comparison between the spatio-temporal dominant approach and the semantics dominant approach.

6.4.3 Reference Spots for Positive and Negative Trajectories

Positive trajectories having periodic visits to medical centres are of interest in this chapter. Thus, only reference spots for the red positive trajectory shown in Figure 6.2(a) is analysed here as an example. In this chapter, we use a time interval of 10 seconds to interpolate irregularly sampled raw trajectories for the spatio-temporal dominant approach. Table 6.2 shows that the semantic dominant approach can obtain more reference spots than the spatio-temporal dominant approach for the positive trajectory under study.

6.4.4 Medical Periodic Patterns for Positive Trajectories

In this section, we present medical periodic patterns using both algorithms, and attempt to infer movement behaviors. Although the spatio-temporal dominant approach does not take background semantic information into

TABLE 6.2: Number of reference spots for the positive trajectory in red shown in Figure 6.2(a).

Approach	Number of reference spots
Spatio-temporal dominant approach	9
Semantic dominant approach	16

account in the process of reference spot detection, we can post-match detected reference spots to nearest medical centres. Figure 6.4(a) displays 9 reference spots for the positive trajectory shown in Figure 6.2(a) whilst Figure 6.4(b) displays 10 reference spots for the negative trajectory using the spatio-temporal dominant approach. The 9 reference spots for the positive trajectory contain medical centres exhibiting frequent visits to medical centres whilst the 10 reference spots for the negative trajectory do not intersect with medical centres.

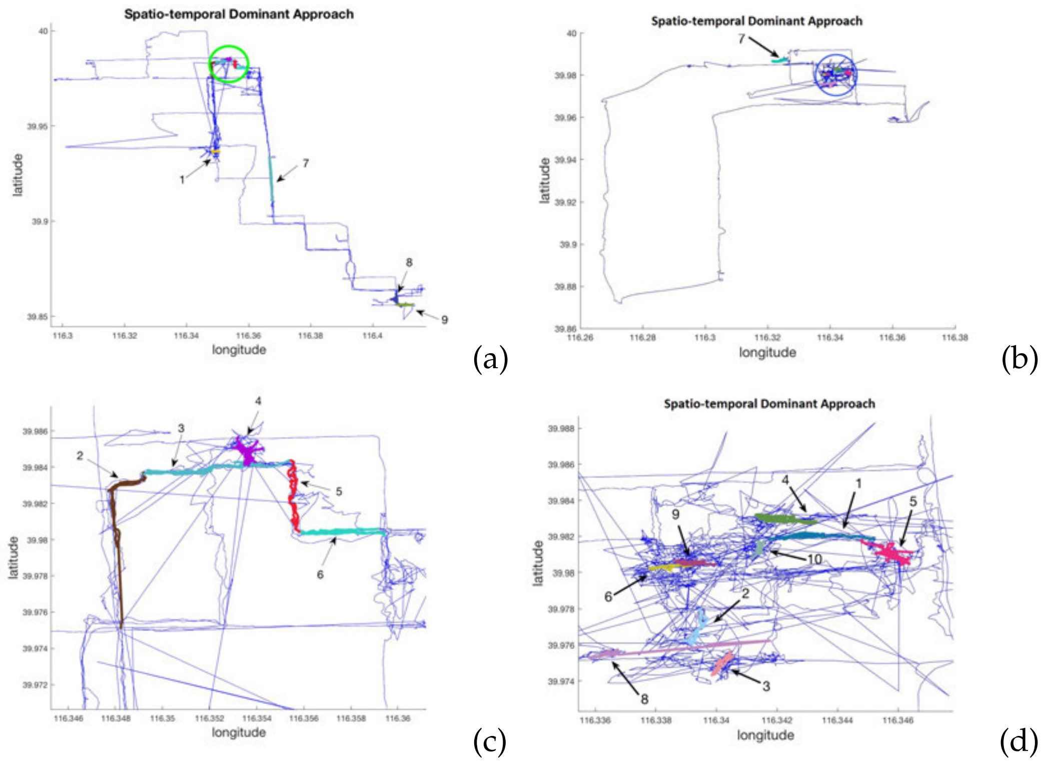
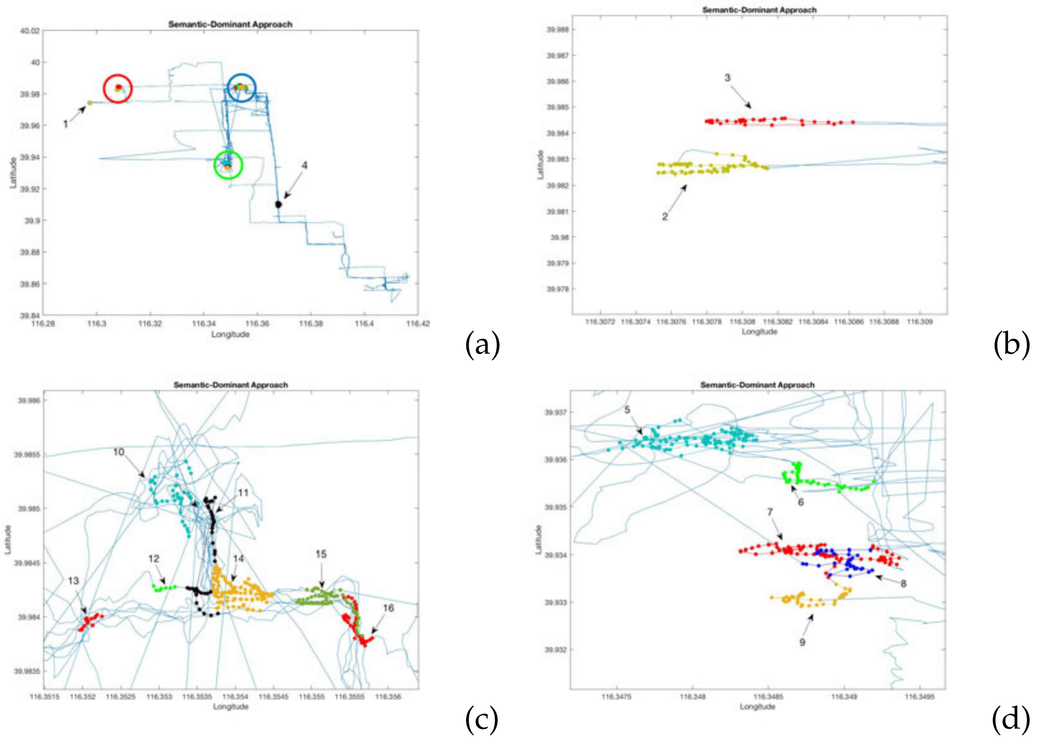


FIGURE 6.4: Obtained reference spots for the positive trajectory and negative trajectory shown in Figure 6.2 Using the spatio-temporal dominant approach: (a) The positive trajectory; (b) The negative trajectory; (c) A zoomed area for (2,3,4,5,6) the green circle in (a); (d) A zoomed area for the blue circle in (b).



(a) (b) (c) (d)

FIGURE 6.5: Obtained reference spots for the positive trajectory shown in Figure 6.2(a): (a) Using the semantic dominant approach; (b) A zoomed area for the red circle; (c) A zoomed area for the blue circle; (d) A zoomed area for the green circle.

Figure 6.5 shows obtained reference spots for the positive trajectory shown in Figure 6.2(a) using the semantic dominant approach. The arrows and numbers indicate *i*-th reference spots. Figure 6.5(b), Figure 6.5(c) and Figure 6.5(d) show zoomed areas for the red circle, blue circle and green circle in Figure 6.5(a), respectively.

TABLE 6.3: Periodic patterns for the positive trajectory using the spatio-temporal dominant approach.

Reference spot	Period (Hours)	Periodic patterns
8	2	9 → 0 → 8
7	9	6 → 0 → 7

Table 6.3 shows identified periodic patterns for the positive trajectory shown in Figure 6.2(a) using the spatio-temporal dominant approach. As mentioned earlier, this method fails to take background semantic information into account, thus detected reference spots are not necessarily matched with medical centres. In this case, two periodic patterns detected are shown in Table 6.3. Since the spatio-temporal dominant approach focuses on finding periodic paths, reference spots 6-9 do not necessarily

match with the medical centres shown in Figure 6.2(c). For the periodic pattern 9 → 0 → 8 and 6 → 0 → 7, reference spots 6-9 are parts of roads as shown in Figure 6.4. Thus, the spatio-temporal approach is not well suited for medical PPM for our study.

TABLE 6.4: Periodic patterns for the positive trajectory using the semantic dominant approach (PU: Peking University; SD: Student Domitory).

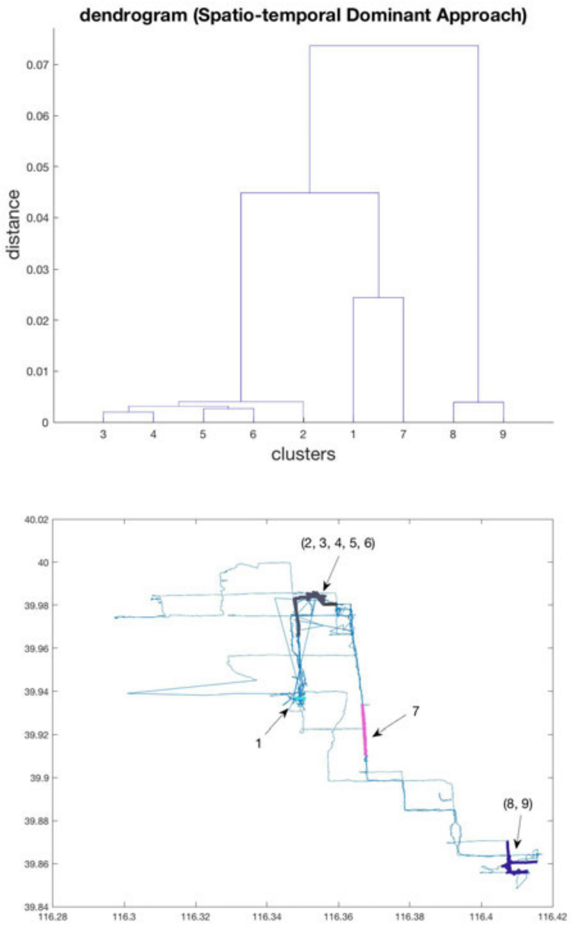
Reference spot	Period (Hours)	Periodic patterns
Building 5	8	PU Hospital 6→ 0 → Building 5
Building 15	3	0 → Building 15
Medical centre 14	23	SD → 0 → Medical centre 14

Table 6.4 shows that obtained periodic patterns using the semantic dominant approach. In Table 6.4, a medical pattern, Peking University People's Hospital 6 → 0 → Building 5, shows a periodic pattern from reference spot 6 (Peking University People's Hospital) to reference spot 5 (Building 5). Reference spot 5 (a building) has a period of 8 hours, which means the user goes to reference spot 5 (Building 5) every 8 hours. 0 means the moving object is not in any reference spot. Another periodic pattern is reference spot 13 → 0 → reference spot 14, where reference spot 14 is matched to a medical centre whilst reference spot 13 is a student dormitory. This medical periodic pattern shows the user goes to the medical centre from the student dormitory periodically with a period of 23 hours. In addition, 0 → reference spot 15 presents a periodic pattern for reference spot 15 (a building) with a period of 3 hours.

Based on these periodic patterns, we can infer this user's health-related movement behaviors. There are two possible inferences: (Peking University People's Hospital 6→ 0 → Building 5) this user needs a periodic medical treatment at Peking University People's Hospital and goes to Building 5 at a period of 8 hours. The user might have a treatment for few hours at hospital and comes back to Building 5 for resting; (Student dormitory → 0 → Medical centre 14) this person could be a student living in a student's dormitory, he/she needs to go to a medical centre regularly at a period of 23 hours. The student might need to have a light treatment everyday at the medical centre.

To sum up, the semantic dominant approach is able to classify a user's movement (trajectory) into a positive trajectory or a negative trajectory, and also it finds medical periodic patterns for the positive trajectory. These medical periodic patterns could be used for hypothesis generation or further inference analysis.

6.4.5 Hierarchical Medical Periodic Patterns for Positive Trajectories



(a)

(b)

FIGURE 6.6: Obtained dendrogram and hierarchical reference spots for the positive trajectory shown in Figure 6.2(a) using the spatio-temporal dominant approach: (a) Dendrogram; (b) Hierarchical reference spots.

Figure 6.6 displays hierarchical reference spots obtained from the spatio-temporal dominant approach. Figure 6.6(a) shows an obtained dendrogram which illustrates the hierarchical relationship between reference spots. Such as reference spot 2, 3, 4, 5 and 6 can be merged as a reference spot (2, 3, 4, 5, 6).

TABLE 6.5: Hierarchical periodic patterns for the positive trajectory using the spatio-temporal dominant approach.

Reference spot	Period (Hours)	Periodic patterns
(8, 9)	6	7 → 0 → (8, 9)
(2, 3, 4, 5, 6)	4	1 → 0 → (2, 3, 4, 5, 6)

Table 6.5 shows identified hierarchical periodic patterns for the positive trajectory shown in Figure 6.2(a) using the spatio-temporal dominant approach. As mentioned earlier, this method fails to take background semantic information into account, thus detected reference spots are not necessarily matched with medical centres. In this case, two hierarchical periodic patterns detected are shown in Table 6.5. Since the spatio-temporal dominant approach focuses on finding periodic paths, similarly with single-level reference spots, two hierarchical reference spots (2, 3, 4, 5, 6) and (8, 9) still do not match with the medical centres shown in Figure 6.2(c). For these two periodic patterns, two hierarchical reference spots are still parts of roads as shown in Figure 6.6(b). Thus, the hierarchical spatio-temporal approach is not well suited for medical PPM for our study.

Figure 6.7(a) shows an obtained dendrogram for the semantic-dominant approach.

TABLE 6.6: Hierarchical periodic patterns for the positive trajectory using the semantic dominant approach.

Reference spot	Period (Hours)	Periodic patterns
(15, 16)	6	14→ 0 → (15, 16)
(10, 11, 12, 14)	12	13→ 0 → (10, 11, 12, 14)
(10, 11, 12, 13, 14, 15, 16)	12	(5, 6) → 0 → (10, 11, 12, 13, 14, 15, 16)
(5, 6, 7, 8, 9)	8	(10, 11, 12, 13, 14, 15, 16) → 0 → (5, 6, 7, 8, 9)

Table 6.6 shows some meaningful multi-level periodic patterns that can match with certain medical centres. A hierarchical reference spot (15, 16) is comprised of a conference room (sometimes, it is used for informal conference), a supermarket and a restaurant. We can call this area non-working area. A single reference spot 14 is one of the medical centres in Peking University Health Science Centre. A hierarchical reference spot (10, 11, 12, 14) includes student dormitories, teaching building and medical centre 14, we can call this area the main activity area. A hierarchical reference spot (10, 11, 12, 13, 14, 15, 16) can be called the whole Peking

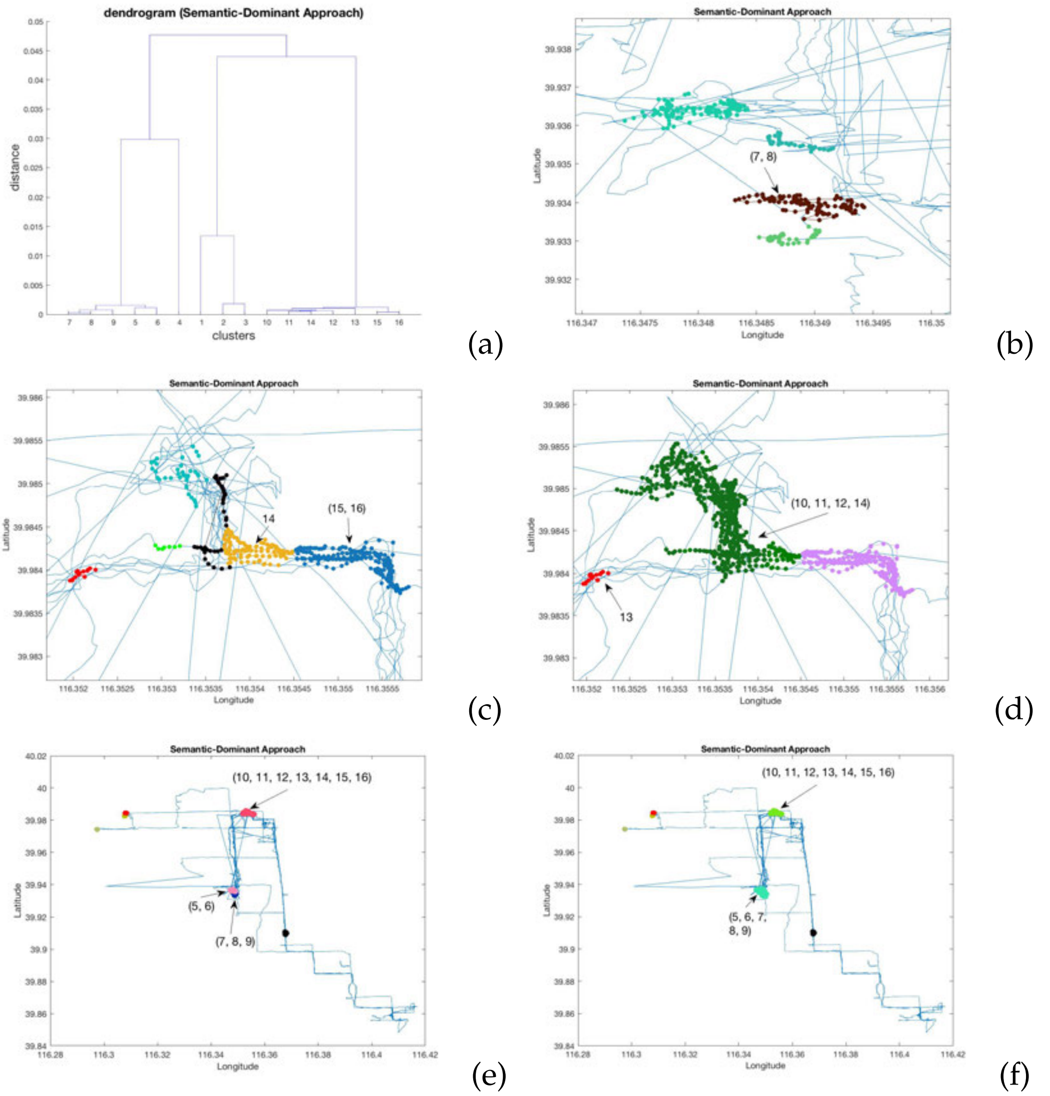


FIGURE 6.7: Obtained dendrogram and hierarchical reference spots for the positive trajectory shown in Figure 6.2(a) using the semantic-dominant approach: (a) An obtained dendrogram; (b) - (f) : Hierarchical reference spots.

University Health Science Centre. A single reference spot 6 represents the main building in Beijing People's Hospital whilst a hierarchical reference spot (5, 6) is still a part of Beijing People's Hospital. A hierarchical reference spot (5, 6, 7, 8, 9) can be called the whole Beijing People's Hospital, including inpatient department. A multi-level periodic pattern 14→ 0 → (15, 16) might show the user went to (15, 16) for food, shopping or informal meeting from spot 14 with a period of 6 hours. In periodic pattern 13→ 0 → (10, 11, 12, 14), a single reference spot 13 is a laboratory, this is very normal that the user has a repeating behavior among teaching building, laboratory and medical centre, which shows the user needs to have a repeating activity among teaching building, laboratory and medical centre with a period of 12 hours. We can call this area a medical area. A periodic pattern (5, 6) → 0 → (10, 11, 12, 13, 14, 15, 16) and (10, 11, 12, 13, 14, 15, 16) → 0 → (5, 6, 7, 8, 9) shows that the user has repeating activities between Beijing People's Hospital and Peking University Health Science Centre with a period of 12 hours or 8 hours. Note that, we do not show periodic patterns for hierarchical reference spots (7, 8) and (2, 3), because there are no periodic patterns for these two hierarchical reference spots. In addition, we can infer that the user is a student or teacher not a patient by periodic patterns with hierarchical reference spots. Interestingly, we cannot infer this with periodic patterns with single-level reference spots as discussed in last section.

6.5 Summary

A spatio-temporal trajectory captures a user's movements and is a solid candidate for mining medical periodic patterns. In this chapter, we introduce a PPM based framework for medical pattern detection. We utilise two spatio-temporal PPM approaches to find medical periodic patterns, and demonstrate the feasibility and applicability of the proposed framework in medical settings using a real-world spatio-temporal movement trajectory dataset. Experimental results reveal that the proposed method is able to classify a user's trajectory into a positive trajectory (frequently visiting medical centres) or a negative trajectory (not frequently visiting medical centres), and also able to detect medical periodic patterns for the positive trajectory. These detected medical periodic patterns can be used for hypothesis generation, cause-effect analysis, and other data mining processes. More experimental results with diverse datasets would further validate the robustness of our approach. The semantic dominant approach is a solid approach for our purpose, but this could be further optimised by tightly incorporating semantic medical information into the algorithm. In addition, we extend the single-level reference spots to multi-level reference spots for hierarchical PPM. Experimental results demonstrate that our framework is able to distinguish those who periodically visit medical centres from those not, and also find single-level and multi-level medical periodic patterns revealing interesting single and hierarchical medical behaviours.

Chapter 7

Conclusion

In this final chapter, we summarise the entire research and discuss its contributions to existing studies in Section 7.1. We present our future work in Section 7.2.

7.1 Summary of This Research

In this section, we make a brief summary of each chapter of this thesis. Existing approaches for PPM ignore six key features of spatio-temporal trajectories, ① the sequence of trajectory; ② spatial and temporal aspects together; ③ the hierarchy of space; ④ irregular trajectory; ⑤ background semantic information. ⑥ trajectory path. In Chapter 4 and 5, we propose a relative approach to solve these problems. Finally, in chapter 6, we present a case study to illustrate how medical periodic patterns can be applied in a health context.

- Chapter 4: In this chapter, we introduce the initial work of our research, a path-based approach to identify four drawbacks from existing PPM: then follow with the four specific drawbacks: ① its negligence of the sequence of trajectory; ② its failure to jointly consider spatial and temporal aspects together; ③ its failure to consider the hierarchy of space. ④ its failure to consider trajectory path. Thus, we propose a path-based approach Traclus (spatio-temporal) to address these four drawbacks in the process of PPM. First, we use a trajectory clustering method based on Traclus to find reference spots. This method inherently considers the sequential nature of trajectory automatically. Second, apart from spatial property, we take a triple $\langle direction, speed, time \rangle$ into account when finding reference spots. There are two reasons for this: these properties can cover both the explicit and implicit features of trajectory in order to obtain better behavior patterns, and also these properties can show periodicity, such as, a moving object visiting a place regularly or at a specific time, can show similar speed and direction in the same road segment. Third, to obtain hierarchical reference spots, we apply HDBSCAN, a hierarchical version of DBSCAN, which can generate clusters with varying densities. We modify HDBSCAN to cluster line segments in order to generate hierarchical reference spots. Finally, experimental results indicate our approach was able to obtain more meaningful and effective periodic patterns than existing methods in single-level. In multi-level, the hierarchical method was able to obtain extra reference spots and periodic patterns with a little extra time when compared to single-level. In addition, the results of this chapter are periodic patterns among trajectory paths as opposed to trajectory nodes.

- Chapter 5: In this chapter, we propose a different PPM approach to our previous work that considers all the five crucial drawbacks of existing spatio-temporal PPM approaches. First, we find stop episodes which can consider the sequence of trajectory. Second, when we match each stop episode with geographical semantic places, we not only match each stop episode spatially but also temporally. Third, in the path-based approach in Chapter 4, we employ an interpolation method to make the input trajectory regular and then find reference spots. There is no doubt that additional sampling points reduce efficiency in the process of mining. To avoid this, we apply Lomb-Scargle periodogram to find periods for each place from irregular raw trajectories. Fourth, in Chapter 4, we only get regular and repeating behaviours of a moving object, such as reference spot 1 → reference spot 2 over 24 hours. Reference spot 1 and reference spot 2 are not attached to any aspatial semantic background information. Thus, it is hard to infer object's regular and repeating behaviors without semantic background information. we cannot answer the questions like "Why does this moving object go to that location? And what is his/her aim?" Based on this, our approach enriches raw spatio-temporal trajectories with meaningful semantics information (consider aspatial semantic background information) provided by an external aspatial semantics database (OpenStreetMap). Then, we match each semantic episode (stop episode) to real world places. Fifth, the hierarchy of space is considered in the presence of contextual aspatial semantics information. We employ HDBSCAN for this aim. To validate the effectiveness and efficiency, we use two real datasets: that one is with known ground truth to evaluate the effectiveness of our approach, and the other is to explore semantically meaningful and interesting periodic patterns. Experimental results demonstrate that our approach achieves better performance than existing approaches. In addition, it is interpretable and effective when periodic patterns are enriched with semantic information. In this work, we can obtain periodic patterns with place type semantics instead of simple designation or number notation. For example, a periodic pattern, home → university with 24 hours, indicates that a moving object goes to university from home every day. It is much more readable and understandable than A → B with 24 hours.

- Chapter 6: In this chapter, we present a case study of mining medical periodic patterns with real world datasets. We apply our path-based approach and node-based approach to medical domains to find positive and negative medical trajectories, and experimental results demonstrate the capability and applicability of our methods to medical domains.

7.2 Future Work

There are several directions for future work.

1. In PPM from spatio-temporal trajectories, where they exhibit repeating and regularly moving behaviors, the PPM research for spatio-temporal trajectories requires a large volume of trajectory data to reveal more reliable patterns. Thus, further experiments with larger and longer term trajectories could be conducted to evaluate the validity of our approach;

2. It is important to understand periodic patterns, thus a post-processing of periodic patterns using visualisation would improve the understandability of periodic patterns. To the best of our knowledge, there is no systematic visualisation approach proposed in this area. Future work could investigate a scalable visualisation approach to help users understand those detected periodic patterns;

3. PPM is for mining patterns from a long and single trajectory, and it is not designed for multiple trajectories. However, finding periodic patterns from multiple trajectories is of interest in some scenarios, but it involves complex computations. In future work, we need to extend our approaches to handle these multiple trajectories to mine periodic patterns from multiple trajectories. The aim of this is to find associated periodic patterns among different users. For example, this approach could reveal a periodic pattern such as Bob and John share an identical university lecture timetable, and both visit the same restaurant at the same time. It is evident that they have the same repeating and regular moving behaviors. Furthermore, it reveals they may have a specific relationship, as a couple or as friends;

4. This thesis provides interesting experimental results in medical domains to prove the applicability of our approaches. Since our approach is a general purpose framework that could be applied to diverse disciplines. More case studies in diverse disciplines would further validate the usefulness of our approaches. For instance, according to moving object's periodic behaviors, people can access to some specific services, such as if a person regularly goes to real estate agent, some relevant service providers can push some related information to him/her. It is useful for him/her to make decisions for further planning.

Bibliography

Agrawal, Rakesh and Ramakrishnan Srikant (1994). "Fast Algorithms for Mining Association Rules in Large Databases". In: *Proceedings of the 20th International Conference on Very Large Data Bases*. VLDB '94. Morgan Kaufmann Publishers Inc., pp. 487–499. ISBN: 1-55860-153-8.

— (1995). "Mining Sequential Patterns". In: *Proceedings of the Eleventh International Conference on Data Engineering*. ICDE '95. IEEE Computer Society, pp. 3–14. ISBN: 0-8186-6910-1.

Agrawal, Rakesh, Tomasz Imielinski, and Arun Swami (1993). "Mining Association Rules between Sets of Items in Large Databases". In: *Proceedings of the 1993 Acm Sigmod International Conference on Management of Data*, pp. 207–216.

Alvares, Luis Otavio et al. (2007a). "A Model for Enriching Trajectories with Semantic Geographical Information". In: *Proceedings of the 15th Annual ACM International Symposium on Advances in Geographic Information Systems*. GIS '07. Seattle, Washington: ACM, 22:1–22:8. ISBN: 978-1-59593-914-2.

Alvares, Luis Otavio et al. (2007b). "Dynamic Modeling of Trajectory Patterns Using Data Mining and Reverse Engineering". In: *Tutorials, Posters, Panels and Industrial Contributions at the 26th International Conference on Conceptual Modeling - Volume 83*. ER '07. Auckland, New Zealand: Australian Computer Society, Inc., pp. 149–154. ISBN: 978-1-920682-64-4.

Alvares, Luis Otavio et al. (2007c). *Towards Semantic Trajectory Knowledge Discovery*.

Apostolico, Alberto, Manuel Barbares, and Cinzia Pizzi (2011). "Speedup for a Periodic Subgraph Miner". In: *Inf. Process. Lett.* 111.11, pp. 521–523. ISSN: 0020-0190.

Aref, Walid et al. (2004). "Incremental, Online, and Merge Mining of Partial Periodic Patterns in Time-Series Databases". In: *IEEE Transactions on Knowledge and Data Engineering* 16, pp. 332–342.

Ashley, Philip W. Suckling; Walker S. (2006). "Spatial and Temporal Characteristics of Tornado Path Direction". In: *The Professional Geographer* 58 (1), pp. 20–38.

Bar-David Shirli; Bar-David, Israel; Cross Paul C.; Ryan Sadie J.; Knechtel Christiane U.; Getz Wayne M. (2009). "Methods for Assessing Movement Path Recursion with Application to African Buffalo in South Africa". In: *Ecology* 90 (9), pp. 2467–2479.

Berberidis, Christos et al. (2002). "Multiple and Partial Periodicity Mining in Time Series Databases". In: *Proceedings of the 15th European Conference on*

Artificial Intelligence. ECAI'02. Lyon, France: IOS Press, pp. 370–374. ISBN: 978-1-58603-257-9.

Berlingerio, Michele et al. (2007). "Mining Clinical Data with a Temporal Dimension: A Case Study". In: *IEEE 2007 IEEE International Conference on Bioinformatics and Biomedicine*, pp. 429–436.

Bermingham, Luke and Ickjai Lee (2014). "Spatio-temporal Sequential Pattern Mining for Tourism Sciences". In: *International Conference on Computational Science*.

— (2015). "A General Methodology for N-dimensional Trajectory Clustering". In: *Expert Syst. Appl.* 42.21, pp. 7573–7581. ISSN: 0957-4174.

— (2017). "A Framework of Spatio-temporal Trajectory Simplification Methods". In: *International Journal of Geographical Information Science* 31.6, pp. 1128–1153.

Bettini, Claudio et al. (1998). "Discovering Frequent Event Patterns with Multiple Granularities in Time Sequences". In: *IEEE Trans. on Knowl. and Data Eng.* 10.2, pp. 222–237. ISSN: 1041-4347.

Bohn, A. et al. (2003). "Identification of Rhythmic Subsystems in the Circadian Cycle of Crassulacean Acid Metabolism under Thermoperiodic Perturbations". In: *Biological Chemistry* 384 (5), pp. 721–728.

Campello, Ricardo J. G. B., Davoud Moulavi, and Jrg Sander (2013). "Density-Based Clustering Based on Hierarchical Density Estimates." In: *PAKDD (2)*. Ed. by Jian Pei et al. Vol. 7819. Lecture Notes in Computer Science. Springer, pp. 160–172. ISBN: 978-3-642-37456-2.

Cao, Huiping, David W. Cheung, and Nikos Mamoulis (2004). "Discovering Partial Periodic Patterns in Discrete Data Sequences". In: *Advances in Knowledge Discovery and Data Mining*. Ed. by Honghua Dai, Ramakrishnan Srikant, and Chengqi Zhang. Springer Berlin Heidelberg, pp. 653–658.

Cao, Huiping, Nikos Mamoulis, and David W. Cheung (2007). "Discovery of Periodic Patterns in Spatiotemporal Sequences". In: *IEEE Trans. on Knowl. and Data Eng.* 19.4, pp. 453–467. ISSN: 1041-4347.

Chanda, Ashis Kumar et al. (2015). "An Efficient Approach to Mine Flexible Periodic Patterns in Time Series Databases". In: *Eng. Appl. Artif. Intell.* 44.C, pp. 46–63. ISSN: 0952-1976.

Chanda, Ashis Kumar et al. (2017). "A New Framework for Mining Weighted Periodic Patterns in Time Series Databases". In: *Expert Syst. Appl.* 79, pp. 207–224.

Chen, Shih-Sheng, Tony Cheng-Kui Huang, and Zhe-Min Lin (2011). "New and Efficient Knowledge Discovery of Partial Periodic Patterns with Multiple Minimum Supports". In: *J. Syst. Softw.* 84.10, pp. 1638–1651. ISSN: 0164-1212.

Dong, Guozhu (2009). *Sequence Data Mining*. Berlin, Heidelberg: Springer-Verlag. ISBN: 1441943528, 9781441943521.

Douglas David h; Peucker, Thomas k (1973). "Algorithms for the Reduction of the Number of Points Required to Represent a Digitized Line or its

Caricature". In: *Cartographica The International Journal for Geographic Information and Geovisualization* 10 (2), pp. 112–122.

Ester, Martin et al. (1996). "A Density-based Algorithm for Discovering Clusters a Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise". In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. KDD'96. AAAI Press, pp. 226–231.

Froelich, Wojciech and Alicja Wakulicz-Deja (2009). "Mining Temporal Medical Data using Adaptive Fuzzy Cognitive Maps". In: *IEEE 2009 2nd Conference on Human System Interactions*, pp. 16–23. ISBN: 978-1-4244-3959-1.

Garofalakis, Minos N., Rajeev Rastogi, and Kyuseok Shim (1999). "SPIRIT: Sequential Pattern Mining with Regular Expression Constraints". In: *Proceedings of the 25th International Conference on Very Large Data Bases*. VLDB '99. Morgan Kaufmann Publishers Inc., pp. 223–234. ISBN: 1-55860-615-7.

Giannotti, Fosca et al. (2007). "Trajectory Pattern Mining". In: *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '07. San Jose, California, USA: ACM, pp. 330–339. ISBN: 978-1-59593-609-7.

Glynn, E. F., J. Chen, and A. R. Mushegian (2006). "Detecting Periodic Patterns in Unevenly Spaced Gene Expression Time Series using Lomb-Scargle Periodograms". In: *Bioinformatics* 22 (3), pp. 310–316.

Gu, Qihang et al. (2017). "Inferring Venue Visits from GPS Trajectories". In: *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. SIGSPATIAL'17. Redondo Beach, CA, USA: ACM, 81:1–81:4. ISBN: 978-1-4503-5490-5.

Gudmundsson, Joachim and Michael Horton (2017). "Spatio-Temporal Analysis of Team Sports". In: *ACM Comput. Surv.* 50.2, 22:1–22:34. ISSN: 0360-0300.

Gudmundsson, Joachim, Patrick Laube, and Thomas Wolle (2017). "Movement Patterns in Spatio-Temporal Data". In: *Encyclopedia of GIS*. Ed. by Shashi Shekhar, Hui Xiong, and Xun Zhou. Cham: Springer International Publishing, pp. 1362–1370. ISBN: 978-3-319-17885-1.

H.A. Sneath, Peter and Robert R. Sokal (1963). *Numerical Taxonomy. The Principles and Practice of Numerical Classification*. Vol. 12, p. 573.

Haining, Robert (2003). *Spatial Data Analysis: Theory and Practice*, p. 432. ISBN: 0-521-77437-3.

Halder, Sajal, Yongkoo Han, and Young-Koo Lee (2013). "Discovering Periodic Patterns using Supergraph in Dynamic Networks". In: *Research Notes in Information Science (RNIS)* 14, pp. 148–151.

Halder, Sajal, Md. Samiullah, and Young-Koo Lee (2017). "Supergraph based Periodic Pattern Mining in Dynamic Social Networks". In: *Expert Systems with Applications* 72, pp. 430 –442. ISSN: 0957-4174.

Han, J., J. Pei, and X. Yan (2005). "Sequential Pattern Mining by Pattern-Growth: Principles and Extensions*". In: *Foundations and Advances in Data Mining*. Ed. by Wesley Chu and Tsau Young Lin. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 183–220. ISBN: 978-3-540-32393-8. DOI: 10.1007/11362197_8. URL: https://doi.org/10.1007/11362197_8.

Han, Jiawei (2005). *Data Mining: Concepts and Techniques*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. ISBN: 1558609016.

Han, Jiawei, Wan Gong, and Yiwen Yin (1998). "Mining Segment-Wise Periodic Patterns in Time-Related Databases". In: *Proc. Int. Conf. on Knowledge Discovery and Data Mining*, pp. 214–218.

Han, Jiawei et al. (2004). "Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach". In: *Data Min. Knowl. Discov.* 8, pp. 53–87.

Han, Jiawei, Zhenhui Li, and Lu An Tang (2010). "Mining Moving Object, Trajectory and Traffic Data". In: *Proceedings of the 15th International Conference on Database Systems for Advanced Applications - Volume Part II*. DASFAA'10. Tsukuba, Japan: Springer-Verlag, pp. 485–486. ISBN: 3-642-12097-0.

He, Zhen et al. (2008). "Mining Partial Periodic Correlations in Time Series". In: *Knowl. Inf. Syst.* 15.1, pp. 31–54. ISSN: 0219-1377.

Hinneburg, Alexander and Daniel A. Keim (1998). "An Efficient Approach to Clustering in Large Multimedia Databases with Noise". In: *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*. KDD'98. AAAI Press, pp. 58–65.

Hoyoung Jeung ; Hua Lu, ; Sathe Saket; Man Lung Yiu (2014). "Managing Evolving Uncertainty in Trajectory Databases". In: *IEEE Transactions on Knowledge and Data Engineering* 26 (7), pp. 1692–1705.

Huang, Kuo-Yu and Chia-Hui Chang (2004). "Mining Periodic Patterns in Sequence Data". In: Springer Berlin Heidelberg, pp. 401–410.

— (2005). "SMCA: A General Model for Mining Asynchronous Periodic Patterns in Temporal Databases". In: *IEEE Trans. on Knowl. and Data Eng.* 17.6, pp. 774–785. ISSN: 1041-4347.

Hwang, San-Yih et al. (2005). "Mining Mobile Group Patterns: A Trajectory-based Approach". In: *Proceedings of the 9th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*. PAKDD'05. Hanoi, Vietnam: Springer-Verlag, pp. 713–718. ISBN: 3-540-26076-5, 978-3-540-26076-9.

Ilarri Sergio; lllarramendi, Arantza; Mena Eduardo; Sheth Amit (2011). "Semantics in Location-Based Services [Guest editor's introduction]". In: *IEEE Internet Computing* 15 (6), pp. 10–14.

Ilayaraja, M. and T. Meyyappan (2013). "Mining Medical Data to Identify Frequent Diseases using Apriori Algorithm". In: *IEEE 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering*, pp. 194–199. ISBN: 978-1-4673-5845-3,978-1-4673-5843-9,978-1-4673-5844-6,

Jain, Anil K. and Richard C. Dubes (1988). *Algorithms for Clustering Data*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc. ISBN: 0-13-022278-X.

Jeung, Hoyoung et al. (2008). "A Hybrid Prediction Model for Moving Objects". In: *Proceedings of the 2008 IEEE 24th International Conference on*

Data Engineering. ICDE '08. Washington, DC, USA: IEEE Computer Society, pp. 70–79. ISBN: 978-1-4244-1836-7.

Jiawei Han ; Guozhu Dong, ; Yiwen Yin (1999). "Efficient Mining of Partial Periodic Patterns in Time Series Database". In: *Proceedings of the 15th International Conference on Data Engineering*. ICDE '99. IEEE Computer Society, p. 106. ISBN: 0-7695-0071-4.

Jindal, Tanvi et al. (2013). "Spatiotemporal Periodical Pattern Mining in Traffic Data". In: *Proceedings of the 2Nd ACM SIGKDD International Workshop on Urban Computing*. UrbComp '13. Chicago, Illinois: ACM, 11:1–11:8. ISBN: 978-1-4503-2331-4.

Kargupta Hillol; Srivastava, Jaideep; Kamath Chandrika; Goodman Arnold (2005). "On Periodicity Detection and Structural Periodic Similarity". In: *Proceedings of the 2005 SIAM International Conference on Data Mining*. Vol. 10.1137/1.9781611972757, pp. 449–460. ISBN: 978-0-89871-593-4,978-1-61197-275-7.

Keogh, Eamonn, Stefano Lonardi, and Bill 'Yuan chi'Chiu (2002). "Finding Surprising Patterns in a Time Series Database in Linear Time and Space". In: *In In proc. of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM Press, pp. 550–556.

L. Zhu, Y et al. (2012). "Mining Approximate Periodic Pattern in Hydrological Time Series". In: *Journal of Computational Information Systems*. Vol. 9, p. 515.

Lahiri, Mayank and Tanya Y. Berger-Wolf (2008). "Mining Periodic Behavior in Dynamic Social Networks". In: *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*. ICDM '08. IEEE Computer Society, pp. 373–382. ISBN: 978-0-7695-3502-9.

— (2010). "Periodic Subgraph Mining in Dynamic Networks". In: *Knowledge and Information Systems* 24.3, pp. 467–497.

Lee, Anthony J.T. et al. (2009). "Mining Frequent Patterns in Image Databases with 9D-SPA Representation". In: *Journal of Systems and Software* 82.4. Special Issue: Selected papers from the 2008 IEEE Conference on Software Engineering Education and Training (CSEET08), pp. 603 –618. ISSN: 0164-1212.

Lee Ickjai; Cai, Guochen; Lee Kyungmi (2014). "Exploration of Geo-tagged Photos Through Data Mining Approaches". In: *Expert Systems with Applications* 41 (2), pp. 397–405.

Lee, I. and J. Yang (2009). "Common Clustering Algorithms". In: *Comprehensive Chemometrics*. Vol. 2, pp. 577–618. ISBN: 9780444527011.

Lee, Jae-Gil, Jiawei Han, and Kyu-Young Whang (2007). "Trajectory Clustering: A Partition-and-group Framework". In: *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data*. SIGMOD '07. Beijing, China: ACM, pp. 593–604. ISBN: 978-1-59593-686-8.

Legány, Csaba, Sándor Juhász, and Attila Babos (2006). "Cluster Validity Measurement Techniques". In: *Proceedings of the 5th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases*. AIKED'06. Madrid, Spain: World Scientific, Engineering Academy, and Society (WSEAS), pp. 388–393. ISBN: 111-2222-33-9.

Li, Yang et al. (2016). "Knowledge-based Trajectory Completion from Sparse GPS Samples". In: *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. GIS '16. Burlingame, California: ACM, 33:1–33:10. ISBN: 978-1-4503-4589-7.

Li, Yifan, Jiawei Han, and Jiong Yang (2004). "Clustering Moving Objects". In: *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '04. New York, NY, USA: ACM, pp. 617–622. ISBN: 1-58113-888-1.

Li, Zhenhui and Jiawei Han (2014). "Mining Periodicity from Dynamic and Incomplete Spatiotemporal Data". In: *Data Mining and Knowledge Discovery for Big Data: Methodologies, Challenge and Opportunities*. Ed. by Wesley W. Chu. Springer Berlin Heidelberg, pp. 41–81.

Li, Zhenhui et al. (2010a). "Mining Periodic Behaviors for Moving Objects". In: *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '10. ACM, pp. 1099–1108. ISBN: 978-1-4503-0055-1.

Li, Zhenhui et al. (2010b). "Swarm: Mining Relaxed Temporal Moving Object Clusters". In: *Proc. VLDB Endow.* 3.1-2, pp. 723–734. ISSN: 2150-8097.

Li, Zhenhui et al. (2011). "MoveMine: Mining Moving Object Data for Discovery of Animal Movement Patterns". In: *ACM Trans. Intell. Syst. Technol.* 2.4, 37:1–37:32. ISSN: 2157-6904.

Li, Zhenhui et al. (2012). "Mining Periodic Behaviors of Object Movements for Animal and Biological Sustainability Studies". In: *Data Min. Knowl. Discov.* 24.2, pp. 355–386. ISSN: 1384-5810.

Liu, Yunhao et al. (2007). "Mining Frequent Trajectory Patterns for Activity Monitoring Using Radio Frequency Tag Arrays". In: *Proceedings of the Fifth IEEE International Conference on Pervasive Computing and Communications*. PERCOM '07. Washington, DC, USA: IEEE Computer Society, pp. 37–46. ISBN: 0-7695-2787-6.

Lomb, N. R. (1976). "Least-squares Frequency Analysis of Unequally Spaced Data". In: *Astrophysics and Space Science* 39 (2), pp. 447–462.

Lv, Mingqi et al. (2016). "The Discovery of Personally Semantic Places Based on Trajectory Data Mining". In: *Neurocomput.* 173.P3, pp. 1142–1153. ISSN: 0925-2312.

Mamoulis, Nikos et al. (2004). "Mining, Indexing, and Querying Historical Spatiotemporal Data". In: *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '04. Seattle, WA, USA: ACM, pp. 236–245. ISBN: 1-58113-888-1.

Maqbool, Fahad, Shariq Bashir, and Abdul Baig (2006). "E-MAP: Efficiently Mining Asynchronous Periodic Patterns". In: *Int. J. of Computer Science and Network Security* 6, pp. 174–179.

McKinley, S. and M. Levine (1998). "Cubic Spline Interpolation". In: *Modelling and Simulation in Engineering* 45.1, pp. 1049–1060.

Mumby P. J.; Vitolo, R.; Stephenson D. B. (2011). "Temporal Clustering of Tropical Cyclones and its Ecosystem Impacts". In: *Proceedings of the National Academy of Sciences* 108 (43), pp. 626–630.

Nishi, Manziba Akanda et al. (2013). "Effective Periodic Pattern Mining in Time Series Databases". In: *Expert Syst. Appl.* 40.8, pp. 3015–3027. ISSN: 0957-4174.

Osuri Krishna K.; Mohanty, U. C.; Routray A.; Mohapatra M.; Niyogi Dev (2013). "Real-Time Track Prediction of Tropical Cyclones over the North Indian Ocean Using the ARW Model". In: *Journal of Applied Meteorology and Climatology* 52 (11), pp. 2476–2492.

Özden, Banu, Sridhar Ramaswamy, and Abraham Silberschatz (1998). "Cyclic Association Rules". In: *Proceedings of the Fourteenth International Conference on Data Engineering*. ICDE '98. Washington, DC, USA: IEEE Computer Society, pp. 412–421. ISBN: 0-8186-8289-2.

Parthasarathy, S., S. Mehta, and S. Srinivasan (2006). "Robust Periodicity Detection Algorithms". In: *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*. CIKM '06. ACM, pp. 874–875. ISBN: 1-59593-433-2.

Pei, Jian et al. (2001). "PrefixSpan: Mining Sequential Patterns by Prefix-Projected Growth." In: *ICDE*. Ed. by Dimitrios Georgakopoulos and Alexander Buchmann. IEEE Computer Society, pp. 215–224. ISBN: 0-7695-1001-9.

Pei, Jian et al. (2004). "Mining Sequential Patterns by Pattern-Growth: The PrefixSpan Approach". In: *IEEE Trans. on Knowl. and Data Eng.* 16.11, pp. 1424–1440. ISSN: 1041-4347.

Peuquet, Donna J. (2002). *Representations of Space and Time*. Guilford Publications.

Rabiner L.; Juang, B. (1986). "An introduction to Hidden Markov Models". In: *IEEE ASSP Magazine* 3 (1), pp. 4–16.

Rhee, Injong et al. (2011). "On the Levy-walk Nature of Human Mobility". In: *IEEE/ACM Trans. Netw.* 19.3, pp. 630–643. ISSN: 1063-6692.

Rousseeuw, P. J. (1987). "Silhousettes: a Graphical Aid to the Interpolation and Validation of Cluster Analysis". In: *Computational and Applied Mathematics* 20, pp. 53–65.

Ruf, T. (1999). "The Lomb-Scargle Periodogram in Biological Rhythm Research: Analysis of Incomplete and Unequally Spaced Time-Series". In: *Biological Rhythm Research* 30 (2), pp. 178–201.

Scargle, J. D. (1982). "Studies in Astronomical Time Series Analysis. II - Statistical aspects of spectral analysis of unevenly spaced data". In: *The Astrophysical Journal* 263, pp. 835–853.

Scott Rebecca; Biastoch, Arne; Agamboue Pierre D.; Bayer Till; Boussamba Francois L.; Formia Angela; Godley Brendan J.; Mabert Brice D. K.; Manfoumbi Jean C.; Schwarzkopf Franziska U.; Sounguet Guy-Philippe; Wagner Patrick; Witt Matthew J.; Beger Maria (2017). "Spatio-temporal Variation in Ocean Current-driven Hatchling Dispersion: Implications for the World's Largest Leatherback sea Turtle Nesting Region". In: *Diversity and Distributions* (6), pp. 604–614.

Sheng, Chang, Wynne Hsu, and Mong Li Lee (2006). "Mining Dense Periodic Patterns in Time Series Data". In: *Proceedings of the 22Nd*

International Conference on Data Engineering. ICDE '06. IEEE Computer Society, p. 115. ISBN: 0-7695-2570-9.

Sirisha, G.N.V.G., M. Shashi, and G.V. Padma Raju (2013). "Periodic Pattern Mining-Algorithms and Applications". In: *Global Journal of Computer Science and Technology Software and Data Engineering* 13 (13).

Spinsanti, Laura, Michele Berlingerio, and Luca Pappalardo (2013). "Mobility and Geo-social Networks". In: *Mobility Data: Modeling, Management, and Understanding*, pp. 315–333. ISBN: 9781107021716.

Srikant, Ramakrishnan and Rakesh Agrawal (1995). "Mining Sequential Patterns: Generalizations and Performance Improvements". In: *Research Report RJ 9994, IBM Almaden Research.*

Tanbeer, Syed Khairuzzaman et al. (2009). "Efficient Single-pass Frequent Pattern Mining Using a Prefix-tree". In: *Inf. Sci.* 179.5, pp. 559–583. ISSN: 0020-0255.

Tao, Yufei et al. (2004). "Spatio-Temporal Aggregation Using Sketches". In: *Proceedings of the 20th International Conference on Data Engineering*. ICDE '04. Washington, DC, USA: IEEE Computer Society, pp. 214–. ISBN: 0-7695-2065-0.

Van Dongen, H.P.A. et al. (1999). "A Procedure of Multiple Period Searching in Unequally Spaced Time-Series with the Lomb?Scargle Method". In: *Biological Rhythm Research* 30 (2), pp. 149–177.

Verhein, Florian and Sanjay Chawla (2006). "Mining Spatio-temporal Association Rules, Sources, Sinks, Stationary Regions and Thoroughfares in Object Mobility Databases". In: *Proceedings of the 11th International Conference on Database Systems for Advanced Applications*. DASFAA'06. Singapore: Springer-Verlag, pp. 187–201. ISBN: 3-540-33337-1, 978-3-540-33337-1.

Viterbi, A. (1967). "Error Bounds for Convolutional Codes and An Asymptotically Optimum Decoding Algorithm". In: *IEEE Transactions on Information Theory* 13 (2), pp. 260–269.

Worton, B. J. (1989). "Kernel Methods for Estimating the Utilization Distribution in Home-Range Studies". In: *Ecology* 70 (1), pp. 164–168.

Yang, Jiong, Wei Wang, and Philip S. Yu (2000). "Mining Asynchronous Periodic Patterns in Time Series Data". In: KDD '00, pp. 275–279.

— (2001). "Infominer: Mining Surprising Periodic Patterns". In: *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '01. San Francisco, California: ACM, pp. 395–400. ISBN: 1-58113-391-X.

— (2002). "InfoMiner+: Mining Partial Periodic Patterns with Gap Penalties". In: *In Proceedings of the 2nd IEEE International Conference on Data Mining (ICDM02*. IEEE Press, pp. 725–728.

Yang, Kung-Jiuan et al. (2013). "Projection-based Partial Periodic Pattern Mining for Event Sequences". In: *Expert Syst. Appl.* 40.10, pp. 4232–4240. ISSN: 0957-4174.

Yavas, Gökhan et al. (2005). "A Data Mining Approach for Location Prediction in Mobile Environments". In: *Data Knowl. Eng.* 54.2, pp. 121–146. ISSN: 0169-023X.

Yeh, Jieh-Shan and Szu-Chen Lin (2009). "A New Data Structure for Asynchronous Periodic Pattern Mining". In: *Proceedings of the 3rd International Conference on Ubiquitous Information Management and Communication*. ICUIMC '09. Suwon, Korea: ACM, pp. 426–431. ISBN: 978-1-60558-405-8.

Ying, Josh Jia-Ching et al. (2011). "Semantic Trajectory Mining for Location Prediction". In: *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. GIS '11. Chicago, Illinois: ACM, pp. 34–43. ISBN: 978-1-4503-1031-4.

Yuan, Guan et al. (2012). "An Efficient Trajectory-clustering Algorithm based on An Index Tree". In: *Transactions of the Institute of Measurement and Control* 34, pp. 850–861.

Yuan, Jing et al. (2013). "T-Drive: Enhancing Driving Directions with Taxi Drivers' Intelligence". In: *IEEE Trans. on Knowl. and Data Eng.* 25.1, pp. 220–232. ISSN: 1041-4347.

Zaki, Mohammed J. (2001). "SPADE: An Efficient Algorithm for Mining Frequent Sequences". In: *Mach. Learn.* 42.1/2, pp. 31–60. ISSN: 0885-6125.

Zhang, Dongzhi, Kyungmi Lee, and Ickjai Lee (2018). "Hierarchical Trajectory Clustering for Spatio-temporal Periodic Pattern Mining". In: *Expert Syst. Appl.* 92.C, pp. 1–11. ISSN: 0957-4174.

Zhang, Minghua et al. (2007). "Mining Periodic Patterns with Gap Requirement from Sequences". In: *ACM Trans. Knowl. Discov. Data* 1.2. ISSN: 1556-4681.

Zheng, Yu (2015). "Trajectory Data Mining: An Overview". In: *ACM Trans. Intell. Syst. Technol.* 6.3, 29:1–29:41. ISSN: 2157-6904.

Zheng, Yu and Xiaofang Zhou (2011). *Computing with Spatial Trajectories*. 1st. Springer Publishing Company, Incorporated. ISBN: 1461416280, 9781461416289.

Zheng, Yu et al. (2011). "Recommending Friends and Locations Based on Individual Location History". In: *ACM Trans. Web* 5.1, 5:1–5:44. ISSN: 1559-1131.