Differentiating population spatial behaviour using a standard feature set

A Thesis Submitted to the
College of Graduate and Postdoctoral Studies
in Partial Fulfillment of the Requirements
for the degree of Master of Science
in the Department of Computer Science
University of Saskatchewan
Saskatoon

By
Rui Zhang

Permission to Use

In presenting this thesis in partial fulfilment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the Department of Computer Science
176 Thorvaldson Building
110 Science Place
University of Saskatchewan
Saskatoon, Saskatchewan
Canada
S7N 5C9

Or

Dean
College of Graduate and Postdoctoral Studies
University of Saskatchewan
116 Thorvaldson Building, 110 Science Place
Saskatoon, Saskatchewan S7N 5C9
Canada

Abstract

Moving through space, consuming services at locations, transitioning and dwelling are all aspects of spatial behavior that can be recorded with unprecedented ease and accuracy using the GPS and other sensor systems on commodity smartphones. Collection of GPS data is becoming a standard experimental method for studies ranging from public health interventions to studying the browsing behavior of large non-human mammals. However, the millions of records collected in these studies do not lend themselves to traditional geographic analysis. GPS records need to be reduced to a single feature or combination of features, which express the characteristic of interest. While features for spatial behavior characterization have been proposed in different disciplines, it is not always clear which feature should be appropriate for a specific dataset. The substantial effort on subjective selection or design of feature may or may not lead to an insight into GPS datasets. In this thesis we describe a feature set drawn from three different mathematical heritages: buffer area, convex hull and its variations from activity space, fractal dimension of the recorded GPS traces, and entropy rate of individual paths. We analyze these features against six human mobility datasets. We show that the standard feature set could be used to distinguish disparate human mobility patterns while single feature could not distinguish them alone. The feature set can be efficiently applied to most datasets, subject to the assumptions about data quality inherent in the features.

Acknowledgements

My deepest gratitude goes first and foremost to Dr. Kevin Stanley, my supervisor, who always supports and encourages me throughout my study. Without his consistent instructions and insightful advice, this thesis could not have reached its best form.

I thank Dr. Carl Gutwin, Dr. Debajyoti Mondal, and Dr. Ehab Diab for their invaluable feedback and suggestions on my thesis. I also appreciate Dr. Daniel Fuller, Dr. Scott Bell and Dr. Rachel Engler-Stringer for their great help.

I am also greatly indebted to the group members and schoolmates Winchell Qian, William van der Kamp, Tuhin Paul, Naveen Kumar Kambham, Luana Fragoso, Yang Qin, Xiaoyan Li, Tonghao Chen, Lujie Duan and Bo Pu, who make the years at the University of Saskatchewan happy and memorable.

I am grateful to my husband, Lin Wu, who led me here and who will accompany me all the way forward. My love and my thanks go to my daughter, Emma Jialu Wu, who not only brings me happiness, but also the chance to be a good mom.

Last but not least, I would like to thank my parents, Jiyuan Zhang and Ping Zhang, my sister Wei Zhang, my parents in law, Kun Wu and Gongqiu Chen, for their continuous support and love. They always stand behind me no matter how far away we are and always give me the strength to get over difficulties and enjoy my life.

Contents

List of Tables

List of Figures

List of Abbreviations

CH	Convex Hull
CH10	Convex Hull of Ten Locations with Longest Dwell Time
BA	Buffer Area
DIM	Box-counting Fractal Dimension

1 Introduction

1.1 Motivation

When we perform activities day by day, we move through space and leave a trail of data behind us. These trails, like, GPS traces, can be collected in detail with GPS sensors embedded in smartphones [4] or represented by the nearby cell towers roughly [29]. With such traces, researchers are able to discover where individuals have been, who they are approximate to, which routes they take, and what services they are able to get as shown in Figure 1.1. This basic information enables researchers to understand better how individuals occupy space, how they connect with other people, and how they interact with their living environment. Knowledge about individual behaviour gained from analysis could drive the design of buildings and neighborhoods, inform health and social policy, or contribute to building a more friendly and more convenient society.

Classically, human spatial behaviour data are acquired by field study, surveys, and travel diaries [5, 34]. Surveys and travel diaries [74, 86] usually ask participants to answer questions about what places they visit in their daily life. Such data collection methods has many benefits in terms of understanding trips purposes and user's socioeconomic issues. However, there are several limitations. First, such methods limit the scale of studies because deploying surveys and collecting diaries require substantial labour. Second, surveys and diaries both limit the completeness of records of an individual's daily life. The questionnaires are not able to cover every aspect of people's life, and the diaries can be subjective. Third, these methods are based on human memory, which can impose errors such as missing short walking trips.

The shortcomings of classic data acquisition methods could be overcome with the development of automated electronic measurements, in particular GPS traces. With cell towers, and personal devices equipped with GPS sensor such as cell phones and smart watches, dense and accurate GPS traces can be recorded as needed. Cell phones can also provide other sensor data and information. These sensors enable researchers to investigate more complex studies such as transit planning, social interaction, and even the spread of some diseases [93, 88, 3, 36]. For example, in [9, 10, 99, 93], researchers utilized GPS sensor or inertial sensors such as accelerometer, gyroscope, compass to estimate traffic and road conditions and generate a more efficient transit map. Researchers from Geography and Health have long attempted to measure individual accessibility to healthcare services, which directly influence individual health conditions [81, 49, 58]. In particular, Kwan [49] pointed out that the accessibility to services is constrained by people's daily out-of-home-activities.

A common goal of spatial behaviour analysis is to compare populations or individuals. Studies of human

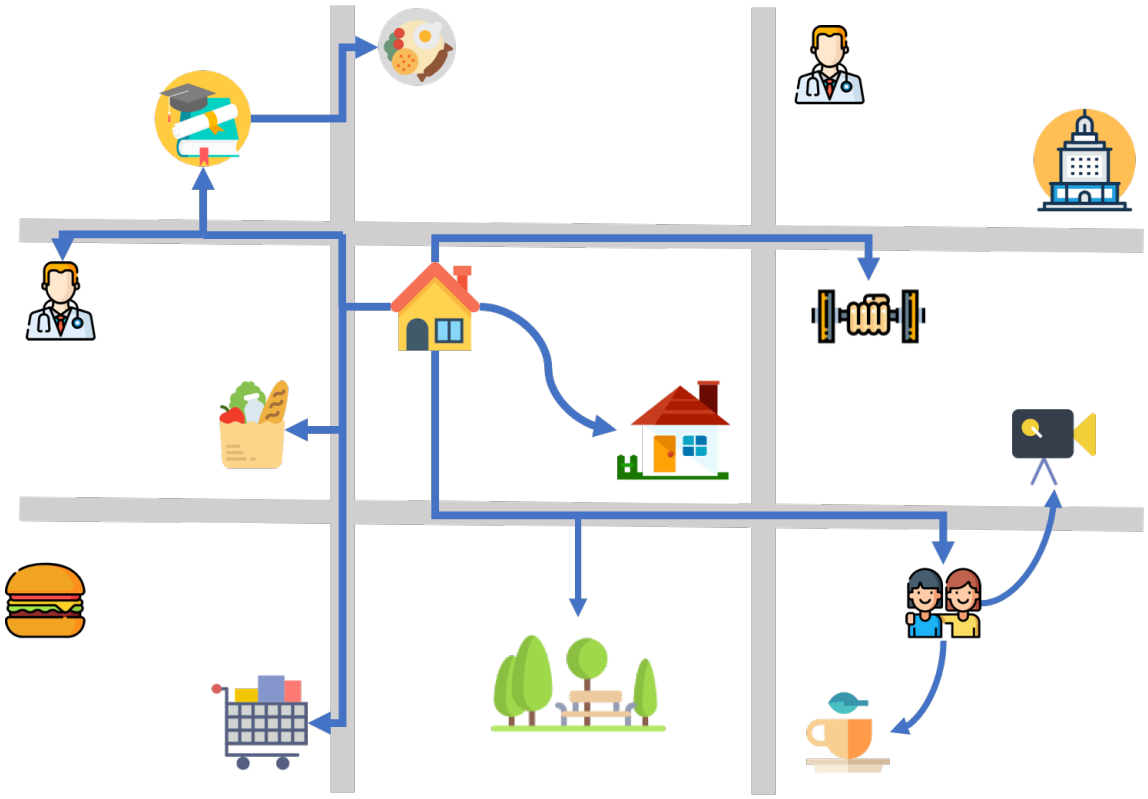

Figure 1.1: Daily activities

spatial behaviour can be broadly categorized into two types: experimental and observational [17, 16]. In observational studies [78, **?**], spatial behaviour is observed and variables are not under control of researchers. Spatial behaviour patterns are extracted from data, and often stratified by demographic parameters such as gender or age group. Researchers manipulate the spatial context explicitly [38] in experimental studies. The natural change of environment [84, 22, 47] is another kind of manipulation in experimental studies. Although GPS traces capture rich details of human spatial movements, it is difficult to compare groups directly. Detailed space time traces should be different from each other, in most cases, eliminating direct comparison as a meaningful outcome. Aggregating or abstracting GPS data to comparable and meaningful representations is necessary for human spatial behaviour studies.

Although a variety of features have been proposed to characterize human mobility, feature design or selection is still a subjective and fraught process. An effective feature should reveal phenomenologically meaningful differences between populations or individuals. Typically, a feature focuses on a specific phenomenon and ignores the others. For example, a classic convex hull polygon of activity space describes the continuous area covered in individual daily movements. However, convex hull method ignores locations inside the polygon and the movement between locations. This trade-off is exacerbated by Modifiable Areal Unit Problem (MAUP) [32], because the results (e.g., area) are influenced by both the shape and scale of the aggregation unit. MAUP leads to different relative values between groups at different levels of spatial

and temporal rate, and the spatial dependency can be difficult to interpret when analyzing a dataset. Scale free or at least scale interpretable features which describe multi-aspects of spatial behaviours would benefit researchers, who need to understand spatial behaviour.

1.2 Problem Statement

Although there are multiple measurements of spatial behaviour from different disciplines and mathematical heritages, there is not a standard feature set for spatial behaviour. In this thesis, we are intended to establish a initial form of a feature set which can describe different aspects of human mobility patterns. Our primary hypothesis is that with the proposed feature set, GPS datasets with significant differences can be distinguished while similar datasets will have similar output.

To evaluate the effectiveness of the proposed feature set, we would validate the feature vectors based on the following fundamental properties with which the features can be useful for a variety of analyses. First, and foremost, features should provide statistical differences between datasets with known differences. If features could not highlight the divergence between datasets, they have limited practical use. Second, features should provide descriptive power to the models that will, or might use them. Some methods, such as convolutional neural network, extract features via layer by layer convolution and down-sampling, and can perform quite well in computer vision tasks. However, such black box features could not clearly show how the features relate to the original data. We expect a descriptive feature set which lead directly to narrative of GPS datasets according to their mathematical properties. Third, features should have strong mathematical foundations which guarantee their valid application. Finally, features should reflect known characteristic of spatial behaviour.

1.3 Contributions

In this thesis, we present the first step towards a standard feature set for human spatial behaviour. We focus on the movements within individual activity space, because individual activity space is related to network design, used mode, activities location, and users preference. We construct the feature set with nine features from different mathematical heritages: buffer area, convex hull and its variations for activity space [25, 16, 76]; fractal dimension, measuring the spatial complexity of GPS traces [20]; and five constant terms of multiscale entropy rate of paths [83, 63, 70], quantifying the predictability of spatial movements.

The proposed feature set was evaluated over six human GPS datasets involving diverse populations from four cities. We present the details of data collection, conditioning, and feature extraction which allow other researchers to validate the feature set on more datasets. Statistical analysis showed that datasets with different demographics or from different geographic areas produced significantly different feature values while datasets which were similar both demographically and geographically generated similar values.

By providing the first detailed description of a standard feature set for analyzing the human GPS mobility traces over their activity space, we have proposed a new tool from extracting to validating feature set for GIS scientists and professionals. The work presented here is not intended to constitute a final form of a standardized feature set, rather it is intended to form the seed of that feature set by drawing features from different mathematical heritages, and provide the process for assessing whether additional or replacement features provide additional benefit, through statistical discrimination of like and unlike datasets.

1.4 Organization of Thesis

This thesis consists of six chapters. Chapter 1 introduces the motivation of the study and the object of this thesis, and points out the general contributions of this thesis. Chapter 2 presents the literature related to the problem and provides strong theoretical background of the features and experiment methods we employ in this thesis. Chapter 3 explains how the experiment is carried out from collecting data, to converting data for feature extraction, and demonstrates how to extract each feature step by step. Chapter 4 analyzes the performance of each feature, the utility of features as a feature set for classification tasks, and the relationship between individual food preference and spatial behaviour. Chapter 5 discusses the shortcomings of the work and the ways to enhance and improve the work in the future. Chapter 6 concludes the overall work and presents the contributions of the thesis.

2 Background

2.1 Literature Review

2.1.1 Collection Methods of Travel Behaviour Data

Researchers have typically collected human spatial data with surveys which include study-related questions. For instance, in a six-week travel diary [5], participants are asked the exact address and purpose of their travel destinations including work, education, daily shopping, etc.. Data collected with surveys or travel diaries could capture the important and accurate places visited in people's daily life. More important, the places mentioned in surveys or dairies are directly related to the type of activities, such as shopping, physical activity, work etc., which can be readily interpreted. But the data may be biased by participants' inaccurate memory and subjective perceptions. Another method is to record participants' behaviours in contrived laboratory studies [35]. However, it's common for participants to behave differently in such environments compared to their normal behaviours. These approaches collect spatial behaviour that people say they perform, but may not be what they actually perform.

There are multiple technological methods to access human locations through time such as Radio Frequency Identification (RFID) [59], social media with geo-located services [75, 27], Automated Fare Collection (AFC) system [71], GSM beacons [61], Call Detail Records (CDR) [29, 43, 8], and Global Position Systems (GPS) [4]. These technologies have enabled researchers to obtain detailed records of human daily movements for a longer term that in a larger population with fewer participant burden than classic methods. Systems that gather, analyze, and present data have been developed for and applied to disciplines such as healthcare, transportation, social networks, and animal science.

In [59], researchers took advantage of the RFID technology used at ticket checkpoints to monitor anonymous users' movement in a large metropolitan mass transit system for a month. Social medias such as Flickr and Twitter which support location metadata provide a digital footprint related to special events or locations [75, 21]. AFC systems have similar application but employ more advanced technology for security and privacy compared to RFID. Data collected with AFC systems or smart cards have been used for improving transit planning [92]. These methods do not record complete human movement, but record individual appearances at specific locations equipped with devices or locations of interest in social media digital footprints.

Among these technological methods, CDR and GPS are the most attractive and widely used. In [29],

researchers study the trajectory of 100,000 anonymized users with their Call Detail Records (CDRs) in a period of six months to understand individual human mobility patterns. Their results indicated that humans follow simple reproducible patterns despite the diversity of their travel history. CDR is capable of tracking a large number of users; however, it is sparse in time and coarse in space which limits its use in characterizing human mobility [8].

Compared to CDR, GPS in general, and smartphone-based GPS in particular, has advantages in providing more varied and finer-grain sources of location information [8], as well as additional data from other federated sensors such as battery, WiFi, and accelerometers. Figure 2.1 shows the most commonly used sensors embedded in smartphones. These sensors capture real-time data which can be used in a variety of fields. The most common applied area of GPS data is transportation. Smartphone sensing platforms such as the MIT VTrack project [89] was designed to provide traffic information using smartphones which could improve the accuracy of travel time estimation and promote commute planning. Further, GPS data is used to analyze activity space environment and health related behaviours such as dietary and physical activities. Zenk *et al.* [97] exploited GPS data over 7 days to examine the relationship among personal demographics, environmental characteristics of neighborhoods, activity space measured with two approaches, and behaviours such as diet and physical activity which affect individual weight. The study proposed that activity space was generally larger than residential neighborhoods and some activity space environmental features were related to dietary behaviours. Accelerometer sensors can capture user's motion, which can further be used to infer different activities (e.g. standing, walking, and running) [52, 53, 95]. A combination of accelerometer, GPS, WiFi and microphones could be used to recognize complex activities including home talking, working, or even emotions [94, 73].

Although both wearable devices and smartphones can capture behavioural data, smartphones are utilized in more studies because participants could use their own smartphones instead of wearing an extra device. Non-sensor data such as battery status, call, SMS logs, and app usage are also available in smartphones. Many applications have been proposed to assist researchers in gathering, analyzing and presenting smartphone data [24, 36, 45]. Eagle *et al.* introduced a smartphone system which collect multiple streams of data from 100 Bluetooth-enabled mobile phones for investigating social networks of individuals and groups. iEpi is a software which was initially designed for epidemiologists and public health researchers [36]. Due to its characteristics of high flexibility, easy reconfiguration, iEpi and its commercial version Ethica were exploited in many pilot studies and experiments on health issues.

2.1.2 Features of Human Spatial behaviour

Human spatial behaviour is complex. It can be difficult to quantify and qualify because it changes over time and interacts with environment. A great deal of efforts have been undertaken to characterize spatial behaviour. There are simple quantitative indicators such as count of unique visited places and complex descriptions about the predictability of movement as potential features available to researchers. In the

Figure 2.1: Smartphone embedded with multiple sensors

following, we reviewed concepts and methods from multiple disciplines which focus on three aspects of spatial behaviour: activity space, complexity of movements, and predictability of spatial behaviour.

Activity space

Activity space is defined as local areas within which people have direct contact in their daily activities [41]. Other similar notions are also widely used in different fields such as action space [40] and potential path area (PPA) [33]. PPA takes the transportation network as well as different constraints such as time restrictions and mobility limitations into account [51], while action space considers individuals' subjective preference or utilization.

As shown in the definition, activity space uses the set of locations with which individuals have interactions in their day-to-day activities. Before the introduction of Geographic Information System (GIS), activity locations such as home, work place and grocery stores were collected using 1 to 2 day dairies. These studies were limited to small group of participants and coarse spatial movement collection. With the GIS technology and widely used smartphones, researchers are able to get almost complete records of daily movements for longer periods over larger populations.

Activity space could be measured by five categories of metrics [66]: minimum convex-hull polygons (MCPs) [25, 16, 76], Standard Deviational Ellipse (SDE) [96] and circles, kernel density estimation (KDE) [6], network-based methods [28]. These methods proposed different assumptions about activity space and result in different shapes (i.e. geographic patterns) and sizes (i.e. coverage).

Minimum convex-hull polygon is the minimum size convex polygon that can surround all points in a data set, as shown in Figure 2.2. In Figure 2.2, the solid red circles in both subplots are the same and represent the locations visited in day-to-day activities. The blue polygon in left subplot depicts the convex hull polygon. As a simple descriptive geometry, this approach yields a continuous area where individuals are able to reach in their daily life. Although convex hull polygons visualize the spatial scope of individuals, it is more likely to be affected by distant locations than clustered points [13].

Standard Deviational Ellipse (SDE) is another widely used metric which captures the orientation and dispersion of spatial points as shown as the yellow ellipse in right subplot in Figure 2.2. Variations of SDE such as circle, Cassini oval, bean curve, and superellipse are also used [74]. SDE works well in situations such as finding spatial patterns of crime [15] where the location data distributes along some geographical objects.

Based on the intuition that people are more likely to visit and have more knowledge about areas near their frequently visited places, Kernel Density Estimation (KDE) is used to spread activity space from the recorded places to the vicinity [50]. In general, the kernel function ensures that locations closer to the central points have larger density than distant locations. We could measure the activity space visualized by KDE with the number of cells whose density is greater than a given threshold. KDE converts visited places into a smooth surface with density as shown in Figure 2.3, but it neglects the connections between locations. This drawback could be concealed for clustered locations, but is magnified for distant isolated locations. For a

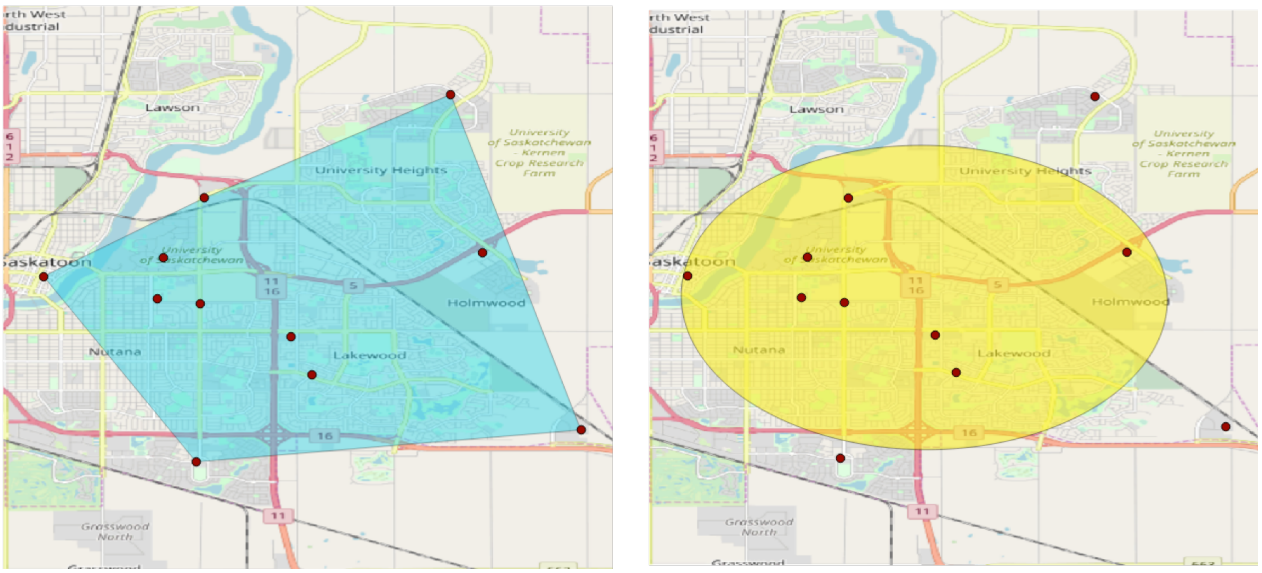

Figure 2.2: Convex hull and standard deviational ellipse of the same datasets

smartphone-based data collection method, this is less of a concern, because such methods collect all visited data points instead of activity locations alone.

All the above approaches suppose continuous space could be utilized by individuals, which is a substantial simplification of real spatial behaviour. Shortest path networks and other network-based methods have been proposed to take the routing network into account as shown in Figure 2.4. This kind of activity space can be measured by length and buffer area of the route generated by GPS traces.

Complexity

While convex hull captures the coverage of individual spatial behaviour, it ignores the detailed activity places and movements between these places which may differ among people. For example, people may visit the same sets of furthest places in their daily life, which leads to same convex hull, but it's almost impossible for them to have exactly same movements at all time. Someone may participate in a variety of activities such as shopping, taking exercise, meeting friends, and so on, while the others prefer to just stay at home except for necessary activities such as driving to work and grocery shopping. The complexity of traces can be captured by fractal dimension.

Fractal geometry was introduced to overcome the limitations of Euclidean geometry when characterizing the irregularity and complexity of objects found in nature, such as coastlines, mountains, and snowflakes. Fractal dimension describes the shape of a complex object, and has been employed in different disciplines successfully, such as ecology [85, 20], medical image analysis [54], and biology [19].

In [20], researchers studied the swimming and searching behaviour in clownfish larvae with conventional analysis as well as fractal dimension analysis. Clownfish usually displays two foraging modes: a linear ranging type of behaviour when searching for food patches and a complex, convoluted behaviour when patches

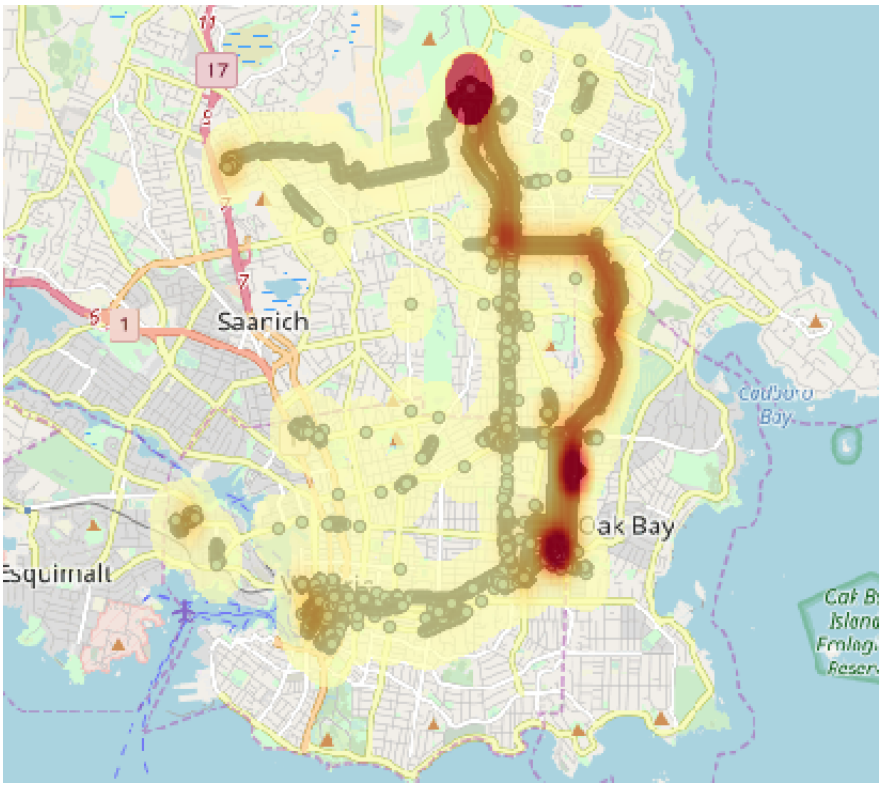

Figure 2.3: Sample of Kernel Density Estimation

are located. Using a three-dimensional tracking of swimming larvae, researchers did not find adequate discrimination between the two modes via conventional analysis including swimming speed and average turning angle. The results of fractal dimension analysis showed that larvae had relatively complex search pattern during the first two days after hatching. When larvae were well fed, the behaviour turned to linear and less complex.

John *et al.* [19] used fractal dimension to measure the complexity in plant development. They calculated the fractal dimension of outlines of 51 fronds from three types of algae. The fractal dimension of mature fronds of all species were not different from each other. The results showed that fractal dimension was correlated with development stage and structural complexity. It revealed that with the development of plants, both size and shape complexity increased.

Although fractal dimension has been applied successfully on measuring the complexity of different shapes including plants, fish movements and others, it has not been used to evaluate human movement traces. In this thesis, we will take fractal dimension as a part of our feature set and verify how individual movement differs from each other from the perspective of complexity.

Predictability

Song [83] first used entropy rate to quantify the variability or predictability of a mobility pattern represented by a string of visited locations. In his study, location of the closest cell tower when individual used his phone

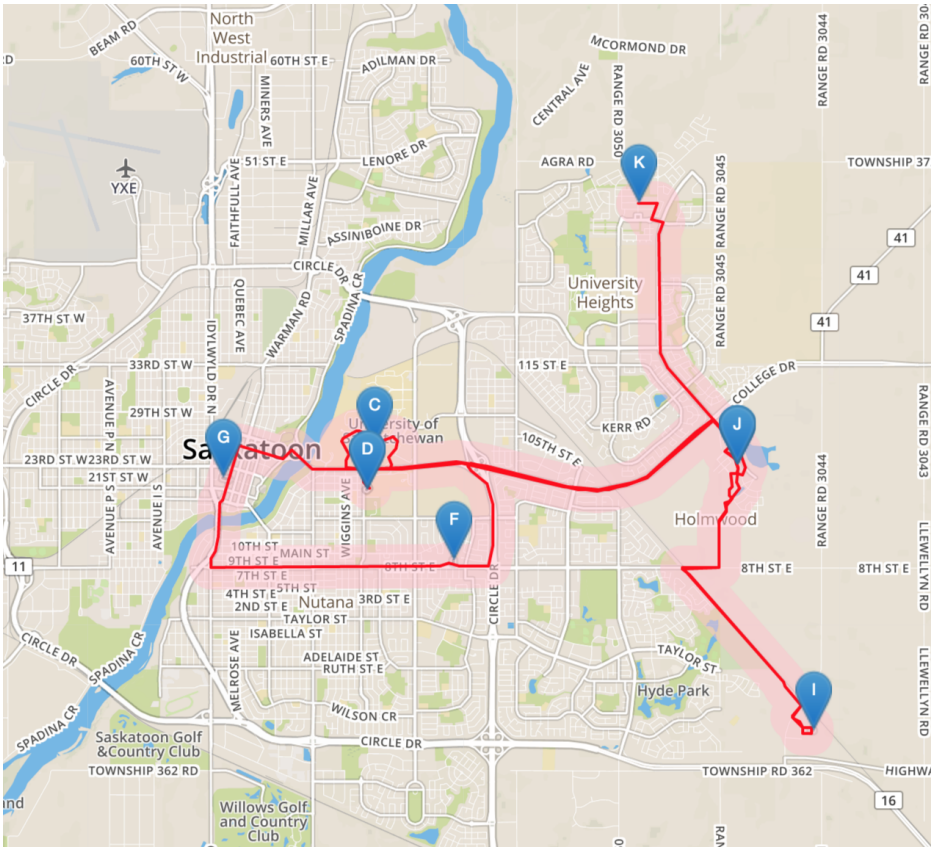

Figure 2.4: Sample of shortest path network method

was treated as the approximate location of mobile users. The spatial temporal trajectory can be converted into a string of locations. The results indicated that a potential of 93% of human behaviour is predictable. Following the approach proposed in [83], researchers [82] calculated the empirical entropy rate using Lempel-Ziv 78 (LZ) algorithm due to its ability of providing asymptotic estimates of entropy rate of a string when its length tends to infinity and a tighter upper bound of mobility predictability was reported which was between 11% - 24% lower than Song's results.

However, mobility entropy rate is not scale invariant as discussed in [72, 82, 63]. Entropy rate calculated using LZ compression algorithm is impacted by varying spatial and temporal sampling. It's problematic to use entropy rate as a comparison metric across datasets which are collected with different spatial and temporal configurations. Osgood *et al.* [63] proposed the scaling relationship of mobility entropy rate and showed the possibility to use entropy rate as a comparison metric among different populations and datasets with different configurations. Paul *et al.* [70] extended Osgood *et al.* [63]'s work and obtained a general scaling relationship that had been validated for varying empirical datasets.

2.2 Methodology

According to above literature review, there are multiple methods which describe various aspects of spatial behaviour. In this thesis, we built a feature set which could distinguish dissimilar datasets and equate alike datasets. To achieve this target, the ideal feature set should be able to summarize spatial behaviour as comprehensively as possible. We select methods for each characteristic, *i.e.*, buffer area, convex hull and its variation for spatial coverage, fractal dimension for complexity of traces, and entropy rate for predictability of movements.

2.2.1 Convex Hull

In mathematics, convex hull is defined as the smallest convex set that contains all points of a given set. As a widely used approach for describing activity space, it produces a minimum convex polygon which covers all locations visited by a participant through his daily movements. Compared to other methods, convex hull doesn't require extra information beyond locations; it is also simple and efficient to compute.

In convex geometry, given a set of points $(x_1, x_2, x_3, ..., x_n)$, a convex combination of the points set is defined as a linear combination of points in the set in the form of Equation 2.1 where coefficients are non-negative and sum to 1 as following:

$$\alpha_1 x_1 + \alpha_2 x_2 + ... + \alpha_n x_n \tag{2.1}$$

If the set only includes two points, convex combinations derived with all possible coefficients lie on the line segment connecting the two points. For a set of two-dimensional points, convex hull can be defined as the set of all convex combinations of points in the set by choosing coefficients in all possible ways. The single

formula for convex hull is:

$$CH(P) = \{\sum_{i=0}^{|P|} \alpha_i x_i | (\forall i : \alpha_i \geq 0) \wedge \sum_{i=0}^{|P|} \alpha_i = 1\} \tag{2.2}$$

Where P is a set of points and α_i is the coefficient assigned to point x_i. Imagine all locations are nails dug into the ground, extend a rubber band to enclose all nails and release it. When the rubber band becomes tense, it reveals the shape of the convex hull.

There are many algorithms for constructing the convex hull of a given set of points or other objects, such as gift wrapping algorithm [18, 44], Graham scan [30], or Quickhull [7]. In this thesis, we use Python package scipy.spatial to compute convex hull, which implements the Quickhull algorithm. The strategy of Quickhull algorithm is to find out the outermost points. The general steps of Quickhull algorithm is shown in Figure 2.5 and described as following.

1. As shown in subplot (a) of Figure 2.5, find the two points A and B drawn in orange in the given set whose distance is the largest. A and B are always vertices of the convex hull polygon. Connect the two points and divide the points set into two subsets drawn in green and blue, respectively.

2. From one side of the dividing line AB, find the point C with the maximum distance to line AB as shown in subplot (b) of Figure 2.5. Remove the points inside the triangle formed by A, B and C, because these points can not be vertices of the convex hull polygon.

3. Repeat the previous step on the line denoted as AC and find the next point D for the convex hull as shown in subplot (c) of Figure 2.5.

4. Repeat the previous two steps until no more points are left. The selected points form the convex hull drawn in blue in subplot (d) of Figure 2.5.

We use the area of convex hull as a feature of spatial behaviour. It reveals the maximum spanning of an individual's movement. A larger convex hull implies that individual is likely or able to move to far locations. That may due to their personality or their transportation mode. A smaller convex hull means a smaller area for activities. This may be true if people live in a well-serviced community.

2.2.2 Convex Hull of Ten Locations with Longest Dwell Time

When we compute convex hull, we consider all locations visited by a participant during a study. This can unnecessarily privilege outliers – places outside routinely visited locations at the edge of activity space. Before the development of GPS, convex hull was constructed from points which were typical activity locations such as home, work, and other routinely visited locations [12]. This classic interpretation avoids the effect of outliers but also limits the complete picture of participant mobility. An alternative formulation of convex hull which balances the outliers and spatial range of participants movements computes the area circumscribed by only those locations frequently visited. Inspired by the study [80] which extracted a mean of 12.7 activity

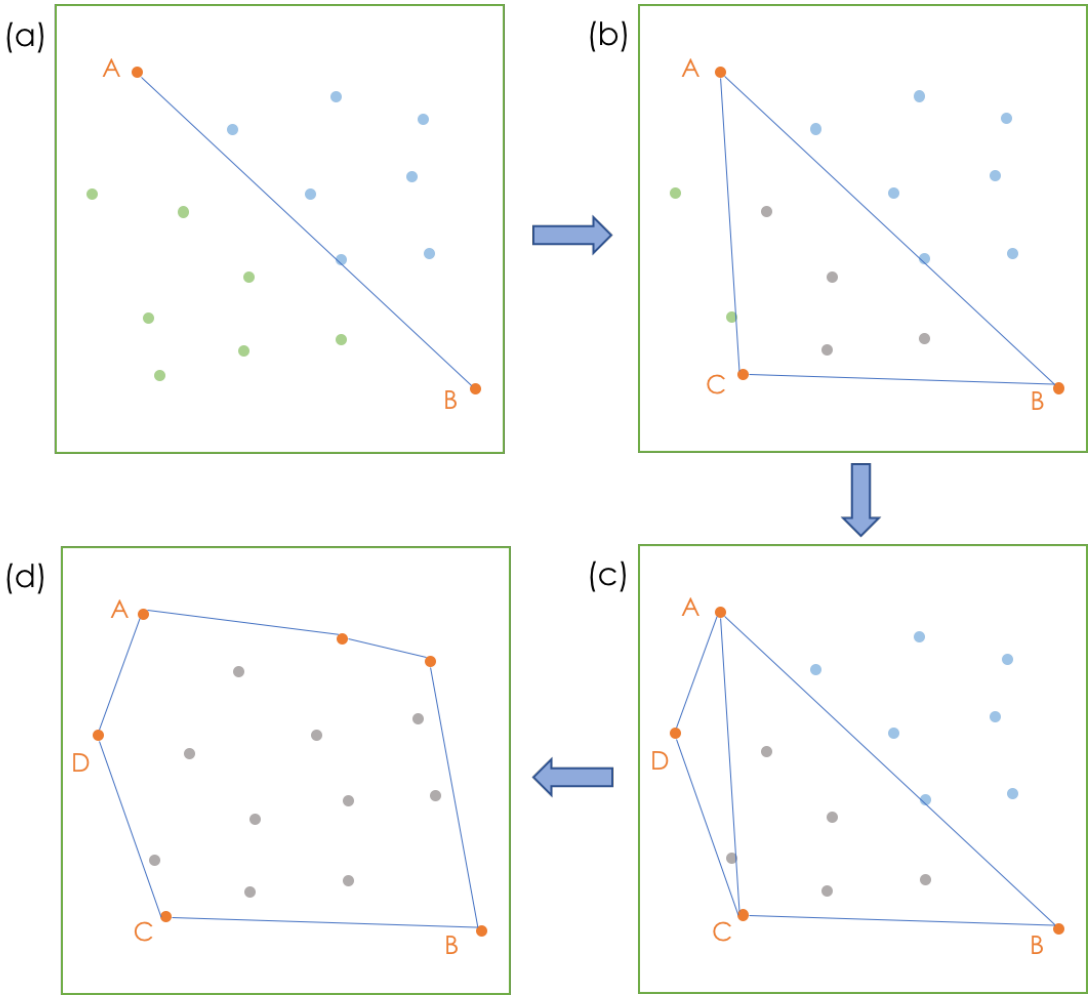

Figure 2.5: Steps of Quickhull algorithm

locations from one-week GPS tracks, we chose 10 places where a participant dwelled for the longest time to construct convex hull.

2.2.3 Buffer Area

As discussed in [77], convex hull method assumes that people make use of continuous space they can reach, which is a simplification of human behaviour considering the limitations imposed by built environment. Based on the notion that areas with which people are familiar are related to their actual travel through space and constrained by transport networks [28], network-based approaches are also widely used to describe activity space. In these methods, activity space is encoded by buffering individual's trips by a (usually fixed) distance as shown in Figure 2.4.

Because GPS locations are not recorded continuously, road networks are often required to create paths between sequence of locations. The most straightforward path is the shortest path [77] according to real road network.

2.2.4 Fractal Dimension

In Euclidean space, a point is described as zero-dimensional, a line is one-dimensional, and a plane is two-dimensional. But some objects are not simple enough to be categorized as a point, a line or a plane such as the Koch snowflake in Figure 2.6. The construction of a Koch snowflake starts with an equilateral triangle. Each line segment is divided into 3 line segments of equal length. Replace the middle segment with an equilateral triangle with the middle segment as its base side, and delete the base side. Repeat the above processes infinitely for each line segment in this shape, then we can get the Koch snowflake. If we zoom the snowflake to 1/6, the local shape is similar to the whole snowflake. And the snowflake is irregular at any scale. It is too complex to be one-dimensional, but is not complex enough to be two-dimensional. The dimension should be some real number between one and two, and is proposed as fractal dimension.

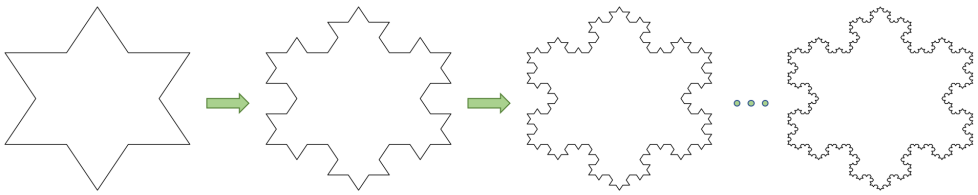

Figure 2.6: The construction of a Koch snowflake

Fractals and fractal dimension were proposed by Benoit Mandelbrot [55, 57, 56]. The fundamental features of a fractal object as addressed are self-similarity and irregularity such as Koch snowflake and coastlines. The perimeter of Koch snowflake varies if we use rulers of different length to measure it because it has details no

matter to what scale it is zoomed.

A fractal dimension describes how complex a fractal is. If the object is a regular shape such as a line, or a circle, fractal dimension would be the same as its dimension in Euclidean space. But fractal dimension of the Koch snowflake will be greater than 1 because it is much more complex than a simple line. Hausdorff [37] introduced the following quantity to define fractal dimension:

$$\Gamma_H^d(r) = \inf_{c_i} \sum_i (r_i)^d$$

where the set S is covered by cells c_i with diameter r_i ($r_i < r$). This definition means we look for a covering set C which minimizes the sum $\Gamma_H^d(r)$. $(r_i)^d$ is the area covered by cell c_i. A cell can be a line segment, a square, a cube, or a hypercube in different dimensional space. The d-dimensional Hausdorff measure is then defined as:

$$\Gamma_H^d = \lim_{r \to 0} \Gamma_H^d(r) \tag{2.3}$$

Hausdorff proved that if d is less than a crucial value D_h, Γ_H^d would be $+\infty$ while it would be 0 if d is greater than D_h. D_h is defined as the Hausdorff dimension of set S. Hausdorff dimension provides the fundamental concept of fractal dimension, but it is difficult to calculate in practice. Variations of fractal dimension based on Hausdorff dimension were proposed to estimate the fractal dimension, for example, box-counting dimension [64], correlation dimension [31], or packing dimension [46].

The fundamental principle of box-counting dimension is intuitive. Given a 2D object, we use boxes of a side length l to cover the object and count the number of boxes $N(l)$ needed, as shown in Figure 2.7. In one dimension, when we use a ruler of length l to measure a curve, the shorter the ruler, the larger the $N(l)$. So $N(l)$ is inversely proportional to l, i.e. $N(l) \propto \frac{1}{l^d}$. In higher dimensions, cubes or hypercubes can be used. Box-counting dimension is defined as:

$$D_b = \lim_{l \to 0} \frac{\ln N(l)}{\ln(\frac{1}{l})} \tag{2.4}$$

The difference between box-counting dimension and Hausdorff dimension is that box-counting dimension uses boxes of the same size to measure the whole object while Hausdorff dimension allows different size of cells. Box-counting dimension is difficult to apply to high-dimension data because the complexity of algorithm increases exponentially with the number of dimension [14].

The correlation dimension was first introduced by Grassberger to measure strange attractors of dynamical systems. Because of its computational simplicity, correlation dimension is a good replacement of box-counting dimension. It takes correlation between points in the data set into account. The correlation could be interpreted as the density of points. Correlation dimension of a set S is defined as [31]:

$$D_c = \lim_{l \to 0} \frac{\ln(C_m(l))}{\ln(l)} \tag{2.5}$$

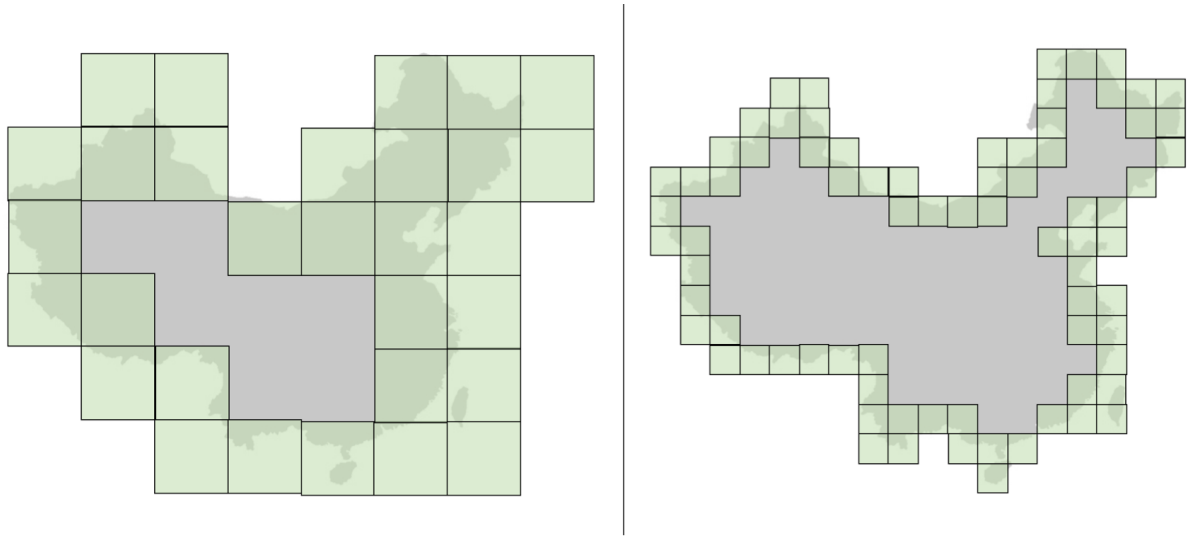

Figure 2.7: Box counting dimension

where $C_m(l)$ is defined as:

$$C_m(l) = \lim_{N \to +\infty} \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} I(\|x_i - x_j\| \leq l) \tag{2.6}$$

I is an indicator function. Value of I is 1 if distance between x_i and x_j is less than l. If not, I has the value of 0. $C_m(l)$ characterizes the proportion of closely related points pairs in the whole data set.

Packing dimension is a modification of box-counting dimension by replacing covering number $N(r)$ with packing number $p(r)$. Information dimension takes the uneven distribution of points into account and adds probability term into dimension definition.

In this thesis, we use box-counting dimension as one feature to characterize the complexity of human movement traces, because we know that human movement is on two dimensional surface.

2.2.5 Entropy Rate

There are some routine trips such as going to work, purchasing grocery at the same store, or taking exercise at the same gym which happen regularly. There are some occasional activities, such as a BBQ at a park which might be spontaneous. The expectation that uncertainty of spatial behaviour should vary among people is not unreasonable [83]. We employ entropy rate to measure the uncertainty or predictability from the other hand.

In information theory, uncertainty is closely related to the probability distribution of a event. If there isn't an asteroid approaching Earth, the probability that Earth will be destroyed by astronomical objects is 0. The event will never happen and there is no uncertainty in this case. If we roll a dice, we will have 6 different outcomes and the probability of getting any number is equal, 1/6. If we roll the dice twice, the size of sample

space will be 36 and the probability of each result is still the same, but lower. As the possible results of an event increase and the probability distribution tends towards uniform distribution, the uncertainty goes up.

As the variables in this thesis are discrete, the probability distribution is represented by probability mass function (PMF), which assigns a value to each outcome in sample space as its probability. Given a random variable $\boldsymbol{\chi}$ and PMF $p(x)$, the uncertainty is represented as entropy and defined in Equation 2.7.

$$H(X) = -\sum_{x \in \chi} p(x) \log_2 p(x) \tag{2.7}$$

Entropy of a variable represents the average uncertainty associated with it. If an event is impossible, entropy is considered to be 0 because $x \log x \to 0$ as $x \to 0$. The upper bound of entropy of χ is achieved when the probability distribution is uniform, which can be easily justified, as shown in Equation 2.8.

$$H(X) \leq |\log_2 \chi| \tag{2.8}$$

Given a sequence of n random variables, it's helpful to know "how does the entropy of sequence changes with n". The rate of growth in a stochastic process $\{X_i\}$ is represented as entropy rate, and defined as following, when the limit exists.

$$H(\chi) = \lim_{n \to +\infty} \frac{1}{n} H(X_1, X_2, ..., X_n) \tag{2.9}$$

For a stationary stochastic process, we can compute the joint PMF $p(x_1, x_2, ..., x_n)$ given a sequence of length N, where $N > n$ [79]. However, the sample space of $(x_1, x_2, ..., x_n)$ grows exponentially with the increase of n which makes it impractical to calculate entropy rate from $p(x_1, x_2, ..., x_n)$. Estimation methods have been proposed to approximate entropy rate. Method based on Lempel-Ziv 78 (LZ) compression algorithm has a faster convergence rate and has been applied to approximate entropy rate of human mobility traces [83]. The LZ-derived entropy rate of a string S of length L as $L \to +\infty$ is defined in Equation 2.10.

$$H = \lim_{L \to +\infty} [(\frac{1}{L} \sum_{i=0}^{L-1} \Lambda_i \ln L)^{-1}] \tag{2.10}$$

where i is the index of character in the string, and Λ_i is the length of the shortest string which starts from i and not found in the substring of S starting from the first character to the character at index $i-1$. Let's take the example given by Tuhin [69] to clarify the process. Given a string "ABABAAAC", where each character represents a distinct location visited by individual. Λ_i for each i starting from 0 is shown in Table 2.1. And $\sum_{i=0}^{L-1} \Lambda_i$ is 16. According to Equation 2.10, LZ-derived entropy rate of "ABABAAAC" equals to $(\frac{16}{8})^{-1} \log_2 8$, *i.e.*, 1.5 bits for base-2 logarithm.

However, as proposed by Osgood *et al.* [63], LZ-derived entropy rate depends on spatio-temporal resolution of the initial discretization of paths. To address this issue, Tuhin *et al.* [70] proposed the following model to separate dependent (path property) and independent (measurement properties) variables of paths by

Table 2.1: Process of LZ_derived entropy rate estimation

Index(i)	Minimum unobserved substring	Λ_i
0	A	1
1	AB	1
2	ABA	3
3	ABAA	3
4	ABAAA	2
5	ABABAAC	3
6	ABABAAAC	2
7	ABABAAAC	1

assuming that a path could be represented by the apparent velocity and dwell time of the agent in each cell. The multiscale entropy rate is expressed as

$$H(d,T) = (d^2\frac{C_1}{4T^2L} + \frac{C_2}{4T^2L} + 2d\frac{C_3}{4T^2L} + d\frac{C_4}{TL} + \frac{C_5}{TL})^{-1}\log L \tag{2.11}$$

where $C_1 = \sum_{i=1}^{n}\frac{1}{v_i^{*2}}$, $C_2 = \sum_{i=1}^{n} t_{d_i}^2$, $C_3 = \sum_{i=1}^{n}\frac{t_{d_i}}{v_i^*}$, $C_4 = \sum_{i=1}^{n}\frac{1}{v_i^*}$, and $C_5 = \sum_{i=1}^{n} t_{d_i}$, where v_i^* is the apparent velocity across the i^{th} cell and t_{d_i} is the total dwell time within the i^{th} cell with side length d. While this formulation provides a mathematically elegant decomposition of entropy rate's dependence on scale and mobility variables, the marginal dwell times and apparent velocities are, by definition, not observable from the location string. To estimate the values of marginal path properties, we follow the technique described in [69] and calculate entropy rate H for a number of downsampled cell sizes d and temporal sampling rates T, and determine a best fit through those points according to Equation 2.11, where the fitted constants are marginal path properties. Finally, we used these five constant terms in Equation 2.11 as features.

2.2.6 Summary of Features

In total, nine features from three disciplines are employed in this thesis. These features are selected for the following reasons:

1. These features commonly exist in any GPS datasets, and not limited to human GPS data. They could be utilized for other datasets such as Moose's and ocean drifters' location data [70]. No extra information is needed for extracting these features.

2. Features are selected considering their distinctive focus on spatial behaviour. We expect that features can capture comprehensive characteristics of spatial behaviour. This will enable researchers to discover the difference between populations.

The features can be interpreted separately as following. The constant terms of multiscale entropy rate are independent features of entropy rate. They do not have explicit meanings as activity space features and fractal dimension. We can interpret these terms by their formulation.

- CH. Convex hull represents the maximum spanning of human movements. Convex hull activity space with larger area implies that an individual is able to reach a larger area. A smaller convex hull may reveal the participant is less likely to go far because of personality or limitation of transition.

- CH10. Convex hull of ten locations with longest dwell time is a conservative activity space measure compared to CH. CH10 of smaller value reveals that frequently visited locations distribute in a small area. Home location is likely to have more impacts on this feature. If people live in a neighborhood which is close to a commercial area or shopping center, they are likely to have CH10 of smaller value.

- BA. Compared to CH and CH10, buffer area is a fixed size area along the daily trip. A larger BA represents a larger travel distance.

- C1 ($\sum_{i=1}^{n} \frac{1}{v_i^{*2}}$). C1 is the sum of one over the squared velocity. The squared velocity ignores direction of movement and is always positive. Quick moves across cells will lead to a smaller C1 compared to slow moves.

- C2 ($\sum_{i=1}^{n} t_{d_i}^2$). C2 is the sum of squared dwell time. It reveals the overall dwelling behaviour of an individual. C2 is larger if people have more and longer dwelling.

- C3 ($\sum_{i=1}^{n} \frac{t_{d_i}}{v_i^*}$). C3 is the sum of quotient of dwell time and velocity. It incorporates dwelling and moving behaviour, which is different from other constant terms. If people dwell a long time and move slowly in a cell, the quotient is large, while a short dwell time and fast move leads to a small quotient.

- C4 ($\sum_{i=1}^{n} \frac{1}{v_i^*}$). C4 is the sum of one of velocity. It is different from C1 because apparent velocity in each cell can be negative considering direction, while squared velocity is always non-negative.

- C5 ($\sum_{i=1}^{n} t_{d_i}$). C5 is the sum of dwell time. Considering that dwell time is each cell is always non-negative, C5 has similar meaning with C2.

- DIM. Fractal dimension (box-counting dimension) considers the entire trip as a polyline and characterize its spatial complexity. If a polyline trip has more details or is less regular, it will get a larger fractal dimension.

There are also shortcomings of these features. All nine features are location independent. Although we can find out the places visited by an individual, we don't know the exact activities people take in these places from GPS data only. This protects individuals' privacy at the cost of a loss at interpretatability. Researchers can compensate this shortcoming with surveys or additional location information such as Google Places.

2.3 Feature Evaluation

To evaluate the discrimination ability of all features, we employed statistical analysis on each single feature and classification model to evaluate the utility of feature combinations. In this chapter, we will introduce

the background of two-step statistical analysis, i.e., ANOVA and Tukey's HSD, and classification model, a support vector machine.

2.3.1 Analysis of Variance

Analysis of Variance (ANOVA) [26] is a widely used statistical hypothesis meant to analyze the differences among group means in a sample. The null hypothesis of ANOVA is that all testing groups are randomly selected from the same population, and means of different groups are equal. For example, in our study, the null hypothesis would be that the spatial behaviour of different populations are all the same. The alternative hypothesis is that different populations do not all exhibit the same spatial behaviour. According to the number of independent variables in experiments, ANOVA can be one-way or two-way. In this thesis, we use one-way ANOVA as the first step to test if the means of six datasets discriminated on each feature.

To perform an ANOVA test, we need to compare two types of variations, i.e., variation between group means and variation within group. The variation between group means is formulated as

$$SS_{between} = \sum n_j(\bar{X}_j - \bar{X})^2 \tag{2.12}$$

where n_j is the number of samples in group j, $\bar{X}$ is the mean of all samples, and $\bar{X}_j$ is the mean of samples in group j. The variations within group is calculated by

$$SS_{within} = \sum \sum (X_i - \bar{X}_j)^2 \tag{2.13}$$

where X_i is a sample in group j. In a test with N samples and k groups, the formula of F value is

$$F = \frac{SS_{between}/k - 1}{SS_{within}/N - k} \tag{2.14}$$

With the calculated F, the null hypothesis can be rejected if F is not less than a critical value. The critical value depends on the level of significance and the degrees of freedom. F statistics has two degrees of freedom, $df_1 = k - 1$ and $df_2 = N - k$. The critical value can be found in a table of probabilities for F distribution with parameters df_1, df_2 and α, where α is the level of significance. Significance level is typically set at 0.05, but can be lower, depending on the application.

2.3.2 Tukey's HSD

Significant results in ANOVA reveal that the group means are not all equal, but it doesn't inform which pair of groups is different. So post hoc analysis is required to further compare all possible pairs of groups. Tukey's HSD (honestly significant difference) test [1], also known as Tukey's range test and other notions, is one of the widely used post hoc tests. Tukey's HSD compares all possible pairs of groups and provides deeper insights into groups.

The Honestly Significant Difference is defined as

$$HSD = q\sqrt{\frac{MSE}{n}} \tag{2.15}$$

where q is the critical value of studentized range statistic, $MSE = SS_{within}/N - k$, and n is the number of samples in each group if group size are equal. Tukey's HSD also works for groups of unequal size as the populations in this study.

For a pair of group (Y_1, Y_2), if $|Y_1 - Y_2| \geq HSD$, Y_1 and Y_2 are significantly different.

2.3.3 Support Vector Machine

With above statistical analysis, we can determine which features distinguish different pairs of populations. To validate the power of features as a whole, we carry out a classification task with Support Vector Machine (SVM) [87].

SVM is an algorithm widely used for classification and regression analysis. The basic idea of SVM is to find a line or a hyperplane in high dimensional space which separates examples of different classes to different sides of the line. SVM can be extended to handle non-linear classification by applying kernels which map an input to high-dimensional feature spaces.

Considering a linear classifier for a binary classification problem with labels $y \in \{-1, 1\}$ and features x. The SVM classifier could be written as

$$h_{w,b}(x) = g(w^T x + b) \tag{2.16}$$

Here, $g(z) = 1$ if $z \geq 0$ and $g(z) = -1$ otherwise. SVM will directly predict either 1 or -1 instead of the intermediate step of estimating the probability of y being 1.

Given a training example $(x^{(i)}, y^{(i)})$, the functional margin of (w, b) is defined as follows:

$$\hat{r}^{(i)} = y^{(i)}(w^T x + b) \tag{2.17}$$

If $y^{(i)} = 1$, we need $w^T x + b$ to be larger to make the functional margin larger. We need $w^T x + b$ to be a large negative number when $y^{(i)} = -1$. A large functional margin means a confident and correct prediction. However, functional margin is not the best measure, because it's sensitive to scale. If we change w to $2w$ and b to $2b$, the prediction won't change given the function g, but the functional margin will be twice as large as the the original value. Thus, we could scale w and b to make the functional margin arbitrarily large without changing anything meaningful. To address this shortcoming of functional margin, SVM uses geometric margin which is a normalization of functional margin. The geometric margin is represented as

$$\hat{r}^{(i)} = y^{(i)}((\frac{w}{\|w\|})^T x^{(i)} + \frac{b}{\|w\|}) \tag{2.18}$$

Geometric margin equals to functional margin if $\|w\| = 1$ and it's invariant to scaling of parameters. Given a training set $S = \{(x^{(i)}, y^{(i)}); i = 1, ..., m\}$, the margin of a hyperplane (w, b) is defined to be the smallest value of geometric margins on the individual training samples:

$$r = \min_{i=1,...,m} r^{(i)} \tag{2.19}$$

We know that given a new point, the larger the geometric margin is, the more confident the prediction will be. To find the separating hyperplane which achieves the maximum geometric margin, the optimization problem is represented as follows:

$$\begin{aligned} & max_{r,w,b} \quad r \\ & s.t. \quad y^{(i)}(w^T x + b) = \hat{r}^{(i)} > \hat{r} \end{aligned} \tag{2.20}$$

The optimization problem could be transformed further based on the feature of functional margin that we can add an arbitrary scaling constraint on w and b without changing anything. The optimization problem after transformation is:

$$\begin{aligned} & min_{r,w,b} \quad \frac{1}{2}\|w\|^2 \\ & s.t. \quad y^{(i)}(w^T x + b) \geq 1, i = 1, ..., m \end{aligned} \tag{2.21}$$

The solution of this nicely transformed optimization problem gives us the optimal margin classifier and could be solved using quadratic programming. The above derivation assumed the data is linearly separable, but there are many data in real world which are not linearly separable. In these cases, SVM exploits different kernels such as polynomial and radial basis function (RBF), to map the original feature x to new features in higher dimensional space in which new features are linearly separable.

3 Experimental Setup

In this thesis, we extracted nine features from GPS locations of participants in several studies, including area of convex hull (CH), area of convex hull of ten locations with longest dwell time (CH10), buffer area (BA), five constant terms ($C1$ to $C5$) of entropy rate considering varying spatial and temporal resolution, and box-counting dimension (DIM). We applied these features to six datasets collected for health or mobility purposes in different populations from three cities in Canada, and a city outside of Canada. All datasets are collected from GPS sensor and/or battery sensor to analyze individual spatial behaviour (although other sensor data such as accelerometer, and phone state information were available in most datasets). Battery data is used to verify the number of potential records available of each participant and to remove participants who did not provide sufficient data. Other post-filtering analysis was based on GPS data. GPS data was preprocessed by filtering out GPS records which are less reliable based on reported accuracy and defined bounding box of each city. The overall process of building a feature set for spatial behaviour is shown in Figure 3.1. The detailed workflow including data filtering, data conversion, feature extraction, and feature validations are shown in Figure 3.2.

3.1 Data Collection

3.1.1 Data Collection Tools

We analyzed six datasets: the Saskatchewan Human Ethology Datasets SHED9 (S9) and SHED10 (S10), a food security dataset (FSD), INTERACT study in Vancouver (VAN) and Victoria (VIC) and taxicabs dataset (TAXI) in Rome. All datasets were collected by devices (mobile phones or tablet) equipped with GPS sensor.

Except for the TAXI dataset, all the other datasets were collected with iEpi [36] (a prototype app) or its commercial successor, Ethica [42]. Both iEpi and Ethica are designed to collect sensor and non-sensor data from off-the-shelf smartphones and to analyze human behaviour based on these data. Typically, GPS and compass/magnetometer are used for navigating, accelerometer and gyroscope provide information about smartphones' motion and rotation. Researchers are able to infer location and activity level from these sensor data. Battery records play an important role in determining individual data quality. If a phone is on, and iEpi/Ethica is running, battery records are guaranteed to be recorded, and are therefore a reasonable measure of data quality, unlike GPS records which may be absent due to circumstances (for example, inside a building, or in an area of poor satellite reception).

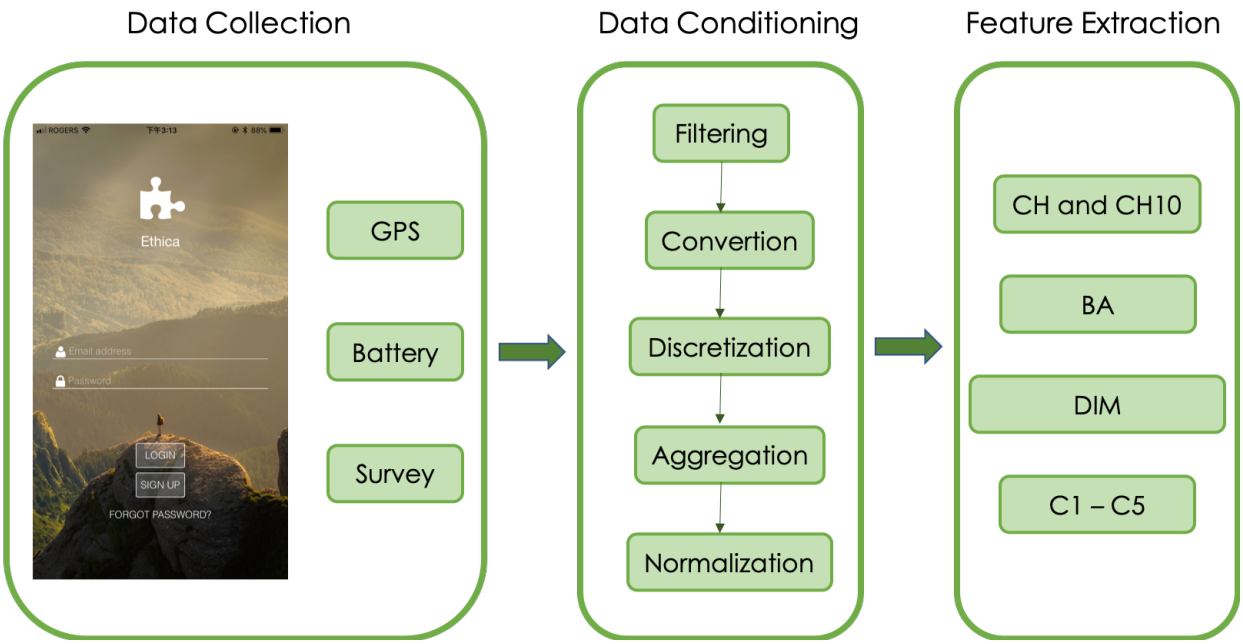

Figure 3.1: The overall process of building a feature set for spatial behaviour

Normally, battery resources of smartphones are depleted quickly if all sensors are activated continuously. iEpi/Ethica collects data on a duty cycle. For example, if the duty cycle is set as five minutes, a smartphone wakes up to collect data for one minute and returns to sleep state during the remaining four minutes, as shown in Figure 3.3. Different duty cycles can be configured for studies based on researchers needs.

Both iEpi and Ethica are able to collect on-phone surveys. Surveys can be triggered regularly (*e.g.*, at 8 AM every day), contextually (*e.g.*, when participants enter a specified area such as campus) or spontaneously (*e.g.*, participants will report grocery shopping as requested). Researchers could also deploy a consent form, demographic survey, and post-survey on Ethica while iEpi uses a paper consent form and a standard Web interface (*e.g.*, surveymonkey) for pre- and post-surveys. All datasets described here were collected with the informed consent of participants under institutional ethics board regulations and approval.

3.1.2 Datasets Introduction

Three of the six datasets were collected from same city, but were distinguished demographically or seasonally. Two datasets were collected as part of the INTERACT study [47] in different cities. The first five datasets all contain data covering the daily lives of participants, collected using a smartphone program, Ethica [42] or its predecessor iEpi [36]. In these datasets, additional sensor modalities (for example accelerometer, gyroscope and WiFi traces) were also collected, but only GPS traces and battery data were used in this study. The last dataset was sourced from a public repository and follows taxicabs [11] rather than individuals. The number of participants, duration and records before and after filter can be found in Table 3.1. The statistics of age and gender of valid participants in each dataset are shown in Table 3.2.

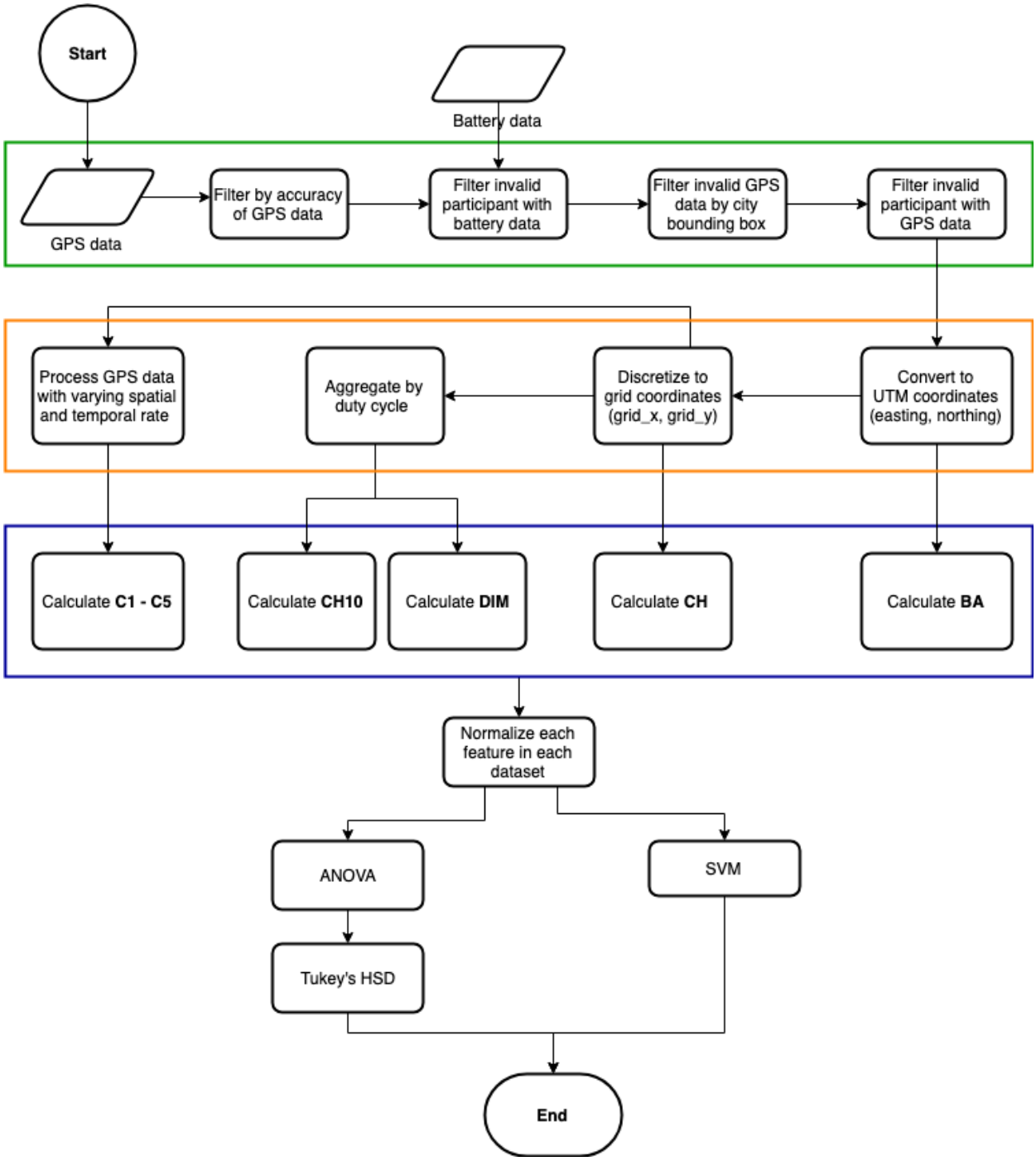

Figure 3.2: The workflow starts from top left. Processes in green rectangle are filtering GPS data step by step. Processes in orange rectangle are converting valid GPS data to different forms for feature extraction. Processes in blue rectangle are extracting features from GPS data in different forms.

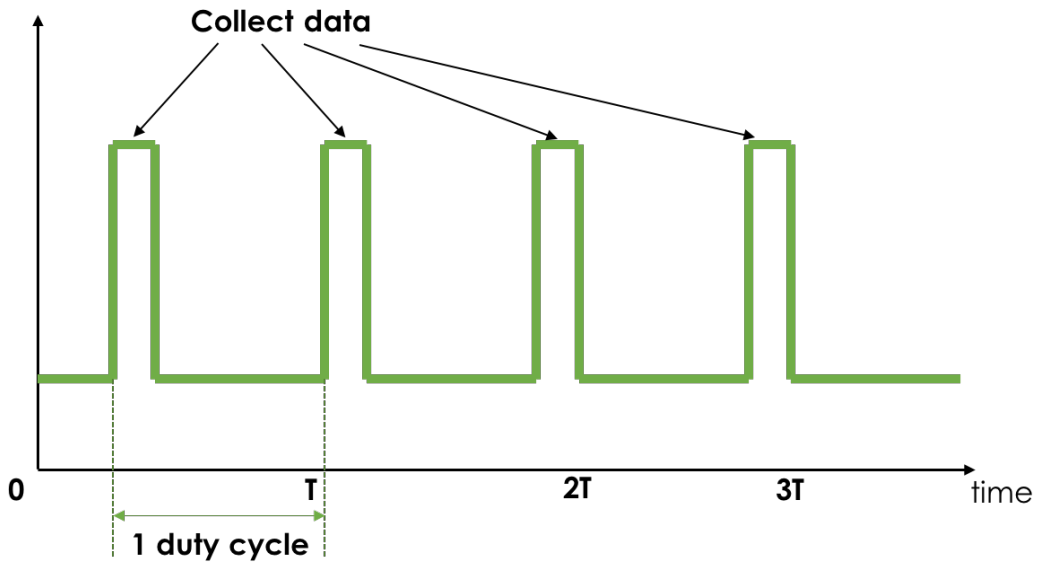

Figure 3.3: The duty cycle mechanism of iEpi/Ethica

Three datasets were collected from the city of Saskatoon, Saskatchewan, Canada. The food security dataset (FSD) was collected as part of a pilot study investigating novel methods for collecting data on how low-income individuals access nutrition. Seventeen low-income families (sometimes with multiple participants) from Saskatoon were involved in the study over a three month period from April to August in 2016 [62].

The Saskatchewan Human Ethology Datasets (SHEDs) are a collection of pilot projects and technical trials related to the development of iEpi, now Ethica, and associated post-processing and methodological outcomes [48, 48, 84]. SHED datasets are exclusively collected from populations at University of Saskatchewan at Saskatoon, Canada. The SHED9 (S9) dataset was collected between October 28, 2016 and December 9, 2016, where 87 students including both undergraduate and graduate students but weighted towards undergraduates were observed. These participants were part of a social science study pool. The SHED10 (S10) dataset was a similar study to S9, where 107 university students drawn from the same social science study pool participated between February 7, 2017 and March 7. Because all three datasets were drawn from the same city, we expect them to exhibit similar spatial scales, but perhaps not extents in activity space. FSD is distinct from S9 and S10 demographically. S9 is similar to S10, with the only notable difference being time of year (Fall versus Winter). S9 and S10 were chosen as a control. We expect them to not be discriminated for most metrics. If metrics distinguish everything, including those things that should be similar, they may be sufficiently sensitive, but insufficiently selective.

Typically, the SHEDs datasets were collected to answer one or more questions about human behaviour. For example, research questions related to food behaviour, physical activity, transportation and stress were running simultaneously on all participants in SHED10, using comprehensive pre/post-surveys composed of groups of questions about each research concern except for general demographic questions. Participants were asked to report their food purchase behaviour and answer several simple questions such as what kind of food

they purchased during the study. We use the food purchase surveys to examine how food purchase preference relates to characteristic of spatial behaviour in this study.

The Vancouver (VAN) and Victoria (VIC) datasets were collected as part of the INTERACT study [47]. INTERACT is a five year, four site, three wave study investigating how changes in urban environment impact health. The Victoria (VIC) dataset is composed of 166 participants who are over 18 years of age and self identified as cycle commuters in the Greater Victoria area of British Columbia, Canada, to study the effect of the implementation of Victoria's All Ages and Abilities (AAA) Bike Network. The Vancouver (VAN) dataset was collected in the Greater Vancouver Area which is the third-largest metropolitan area in Canada. It is designed to reveal how the development of Vancouver's Arbutus Greenway is impacting physical activity, social participation, and well-being of nearby residents, and whether these impacts are felt equally across different socioeconomic groups. A preliminary cohort of 64 participants who are 18 years of age or older and live within 3 km of the Arbutus Greenway were recruited between May 2018 and August 2018. In both the VIC and VAN studies, Ethica smartphone traces were recorded over a one month period. The VIC dataset is distinct in geographic location, as Victoria is a coastal city approximately the same size as Saskatoon, and demographically, as participants are self-identified cycle commuters. We expect Victoria to be similar to Saskatoon in some measures of activity space considering their similar size, but distinct in measures which enhance temporal differences in trajectories. Vancouver is a large metropolitan area, and is expected to be distinct from all other datasets across all measures.

The TAXI dataset [11, 2] was collected with an Android OS tablet device running an app that updated the current GPS position towards a server every 7 seconds. It contains mobility traces of 316 taxi cabs from February 1 to March 2 2014 in Rome, Italy. The TAXI dataset is distinct in many ways from the other datasets: because it tracks taxis, not people, it is expected to have irregular trajectories, and ill-defined activity spaces. Because the taxis are from Rome, the trajectories should be distinct from the Canadian cities in the other datasets.

Table 3.1: Datasets information

Dataset	City	Duration	P (#)	P* (#)	DC rates	Battery (#)	GPS (#)	GPS* (#)
FSD	Saskatoon	90 days	20	10	8 mins	587722	1467785	1014949
S9	Saskatoon	40 days	87	59	5 mins	644367	15878570	12218667
S10	Saskatoon	29 days	107	40	5 mins	353840	8592409	5469949
VIC	Victoria	30 days	166	71	5 mins	1024509	10108481	6606518
25VAN	Vancouver	30 days	64	28	5 mins	341934	1411992	520437
TAXI	Rome	30 days	316	282	7 secs	N/A	21817851	19118793

P: all participants, P*: accepted participants, DC: duty cycle, GPS (#): count of raw GPS records GPS* (#): count of GPS records after filtering.

To ground the discussion on geographic extent and coverage of each dataset, we show the general distribution of GPS records of all datasets with heatmaps in Figure 3.4. The heatmap was drawn by aggregating the filtered GPS records into grids with side length of 250 m. The blue rectangle in each heatmap depicts

Table 3.2: Demographic characteristics of each dataset

Variable		FSD	S9	S10	VIC	VAN
Age	Min	24	18	18	23	36
	Max	63	37	38	67	76
	Mean	35.5	25.1	26.1	42.3	58.1
	Std	13.9	4.8	5.3	11.9	10.6
Gender	Female	8	37	25	31	16
	Male	2	22	15	38	12
	Other	0	0	0	2	0

the defined bounding box of each city. The area within the bounding box of Saskatoon, Victoria, Vancouver, and Rome are 17 km × 17 km, 20 km × 20 km, 75 km × 63 km, and 33 km × 30 km, respectively. The map scales are [FSD: 3 km, S9: 3 km, S10: 3 km, VIC: 5 km, VAN: 10 km, TAXI: 10 km], and are shown at the left bottom of each heatmap. There are more records in the red area than the green area.

3.2 Data Conditioning

3.2.1 Data Description

To explore individual spatial behaviour, locations visited in participants' daily life are important data source. GPS sensor embedded in smartphones record latitude, longitude, altitude, timestamp, accuracy, provider, and speed. Several GPS record samples in the datasets are shown in Table 3.3. Latitude and longitude are coordinates of locations in geographic coordinate system. Accuracy implies how reliable the geographic locations are. Device ID is the unique id for each smartphone. There are two types of provider in GPS table, gps and network-passive. The accuracy of location records from GPS is usually the highest because it uses satellites. When the satellites condition is not good, a network is used to estimate position.

Table 3.3: Example of GPS data

User ID	Device ID	Lat	Lon	Alt	Accu	Record Time	Provider
777	112dde4393bc99b8	52.1312847	-106.6360358	0	21.253	2017-03-01 13:43:25	network-passive
777	112dde4393bc99b8	52.1312759707	-106.6359466146	437.44	12	2017-03-01 13:46:32	gps
777	112dde4393bc99b8	52.1312654685	-106.6359259325	435.981	16	2017-03-01 13:46:33	gps
777	112dde4393bc99b8	52.13127812	-106.6359315539	457.175	12	2017-03-01 13:46:34	gps
777	112dde4393bc99b8	52.1312991674	-106.6359806312	460.254	6	2017-03-01 13:46:35	gps

Lat: latitude, Lon: Longitude, Alt: Altitude, Accu: accuracy

We use battery data to determine data quality at participant level if available. Battery information from

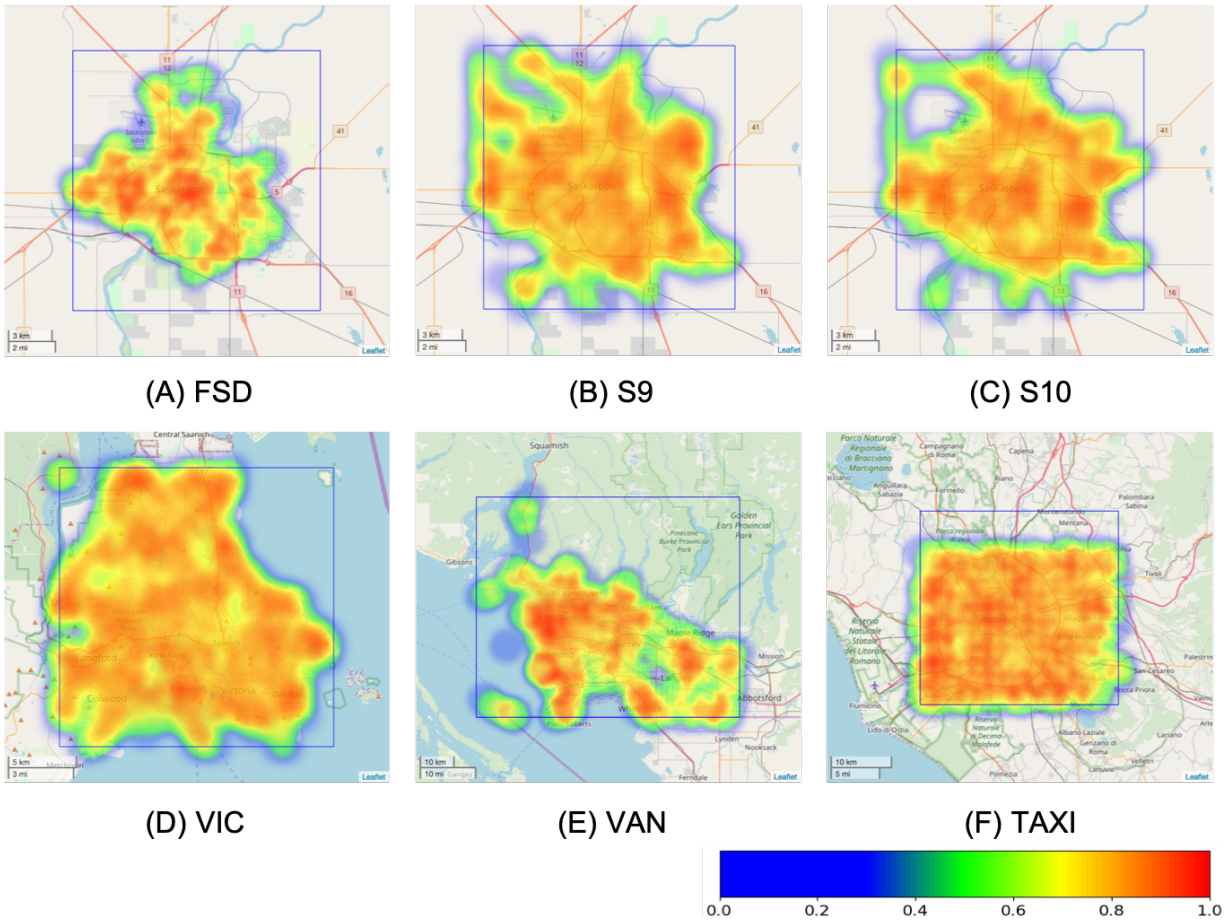

Figure 3.4: Heatmaps of filtered GPS records of each dataset

Ethica includes user ID, device ID, and record time. Samples of battery data are shown in Table 3.4.

Table 3.4: Example of battery data

User ID	Device ID	Record Time	Level	Plugged	Scale	Temperature	Voltage
777	112dde4393bc99b8	2017-02-08 13:26:54	83	0	100	312	4089
777	112dde4393bc99b8	2017-02-08 13:31:54	82	0	100	315	4105
777	112dde4393bc99b8	2017-02-08 13:41:54	80	0	100	217	4069
777	112dde4393bc99b8	2017-02-08 13:46:54	80	0	100	206	4071
777	112dde4393bc99b8	2017-02-08 13:51:54	79	0	100	233	4082

3.2.2 Data Filtering

The complete filtering process includes following steps:

1. To remove erroneous locations, GPS records with accuracy poorer than 100 m were removed.

2. Because we are primarily interested in routine within-city movements and behaviour, GPS traces outside the assumed city bounding box of each city were removed. The defined bounding box of each city is shown in Table 3.5.

3. We remove participants who are less reliable and don't have adequate records for analysis. Participants who had less than half of the maximum possible number of duty cycles containing battery records were excluded from further analysis as shown in Figure 3.5.

4. Participants who have reported less than a quarter of the maximum possible number of duty cycles containing GPS records were excluded as shown in Figure 3.6, because too few records could lead to artifact-driven outliers.

Table 3.5: Range of each city's bounding box

City	Range of latitude	Range of longitude
Saskatoon	(52.058367, 52.214608)	(-106.764914, -106.522253)
Victoria	(48.391892, 48.57277)	(-123.540331, -123.271128)
Vancouver	(49.001407, 49.566829)	(-123.441453, -122.406227)
Rome	(41.769653, 42.052603)	(12.341707, 12.730937)

After filtering out invalid participants and invalid GPS records, the distribution of study duration and daily GPS records count are shown in Figure 3.7 as a description of dataset quality. Because datasets have different duty cycles, we downsampled all datasets to the same duty cycle 40 mins. According to Figure 3.7, S9, S10, VIC, and VAN have similarly high quality, while FSD and TAXI have fewer records.

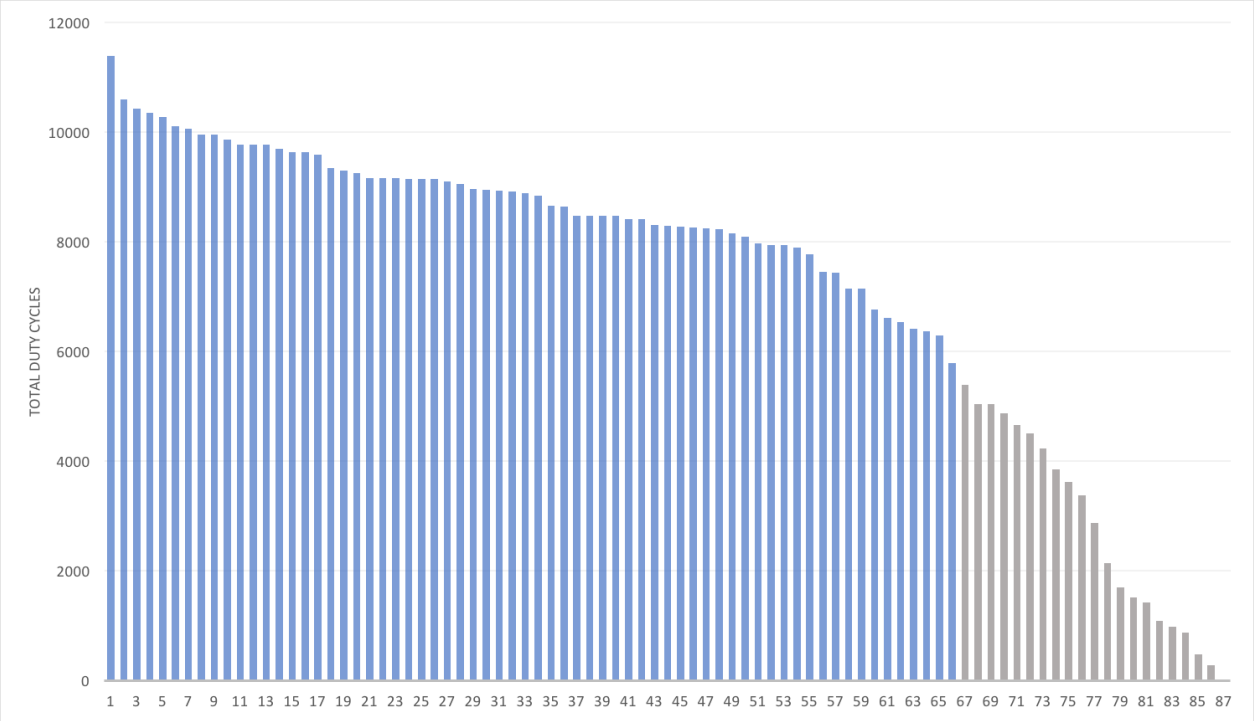

Figure 3.5: X axis is the index of participants in SHED9. Y axis represents the count of all battery duty cycles during the study. Grey bars represent participants who are filtered out because they have less than half of the maximum possible number of duty cycles containing battery records which is 11385.

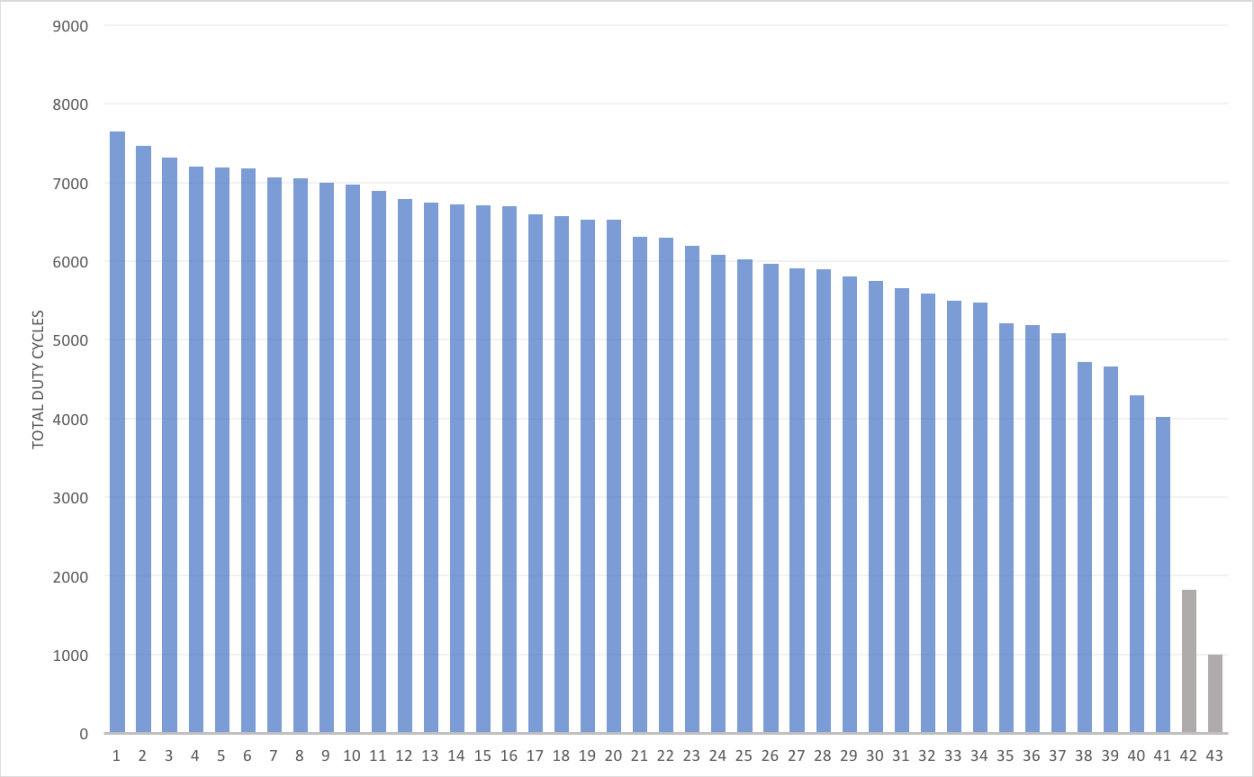

Figure 3.6: X axis is the index of participants in SHED10. Y axis represents the count of GPS duty cycles. Grey bars represent participants who are filtered out because they have less than a quarter of the maximum possible number of duty cycles containing GPS records which is 7652 in SHED10.

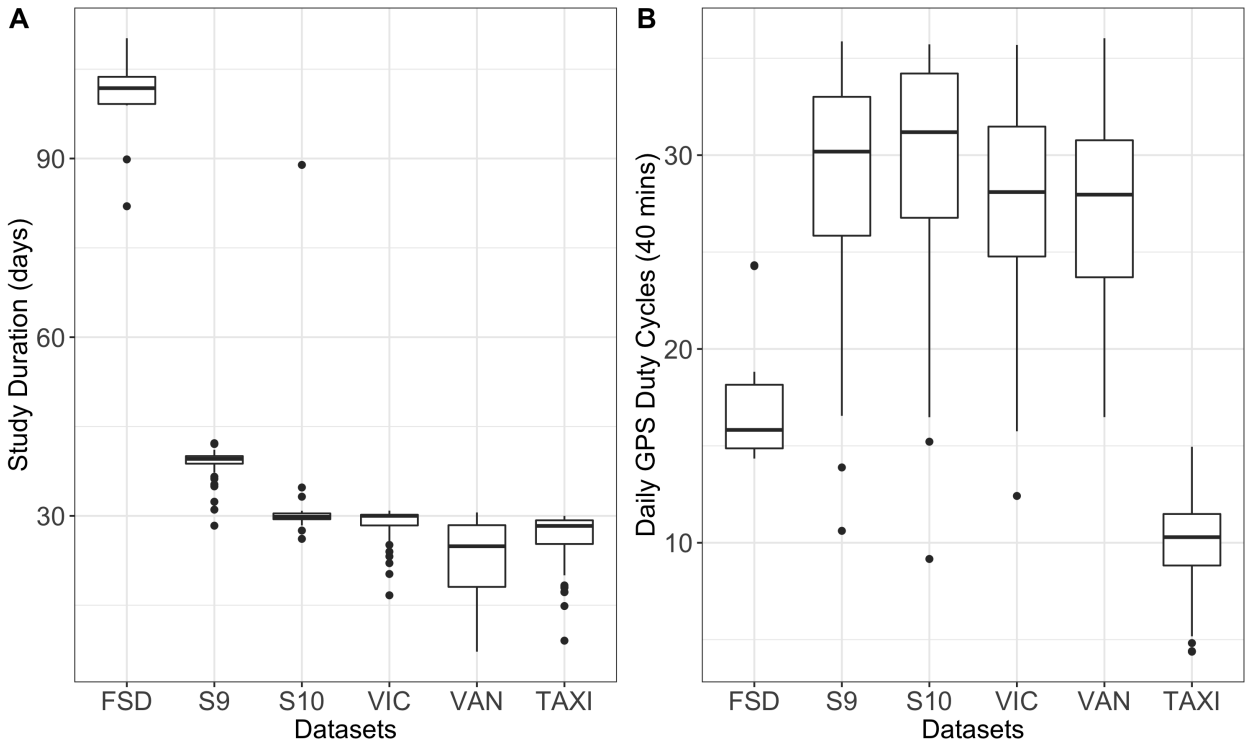

Figure 3.7: Boxplot A illustrates the distribution of study duration of each dataset. Boxplot B shows the daily GPS duty cycles with unit of 40 mins.

3.2.3 Data Conversion

GPS records collected from mobile phones are represented by geographic coordinates, $(latitude, longtitude)$. Geographic coordinate system models the Earth as a 3-dimensional oblate spheroid and defines locations on the spherical surface using the angles measured from the heart of the Earth to locations, that is $(latitude, longtitude)$. Latitudes and longitudes are measured in degrees and represented as decimal numbers. As we focus on individual spatial behaviour, we usually consider geometric measures such as length of trips, area covering all locations.

For computation, we convert geographic coordinates to Universal Transverse Mercator (UTM) coordinates $(easting, northing)$ using Python package pyproj 1.9.5.1. UTM conformal projection models the surface of the Earth as a 2-dimensional Cartesian coordinate system which is ideal for calculating distance, area and other metrics. Because we limit analysis to a single city, all points for each dataset are within a single UTM zone. The EPSG codes used for Saskatoon, Victoria, Vancouver, and Rome are 32613, 32610, 32610, and 32633, respectively.

3.2.4 Data Discretization

To calculate entropy rate, continuous city space needs to be discretized. Following the setting in [69], we used 15.625 m as the base grid resolution. GPS locations were subsequently converted from UTM coordinates to

grid coordinates $(grid_x, grid_y)$. Strategy of the conversion is as following:

$$\begin{aligned} grid_x &= \frac{easting - easting_{min}}{grid\_size}, \\ grid_y &= \frac{northing - northing_{min}}{grid\_size} \end{aligned} \tag{3.1}$$

where $easting_{min}$ and $northing_{min}$ are the minimum easting and minimum northing value of the defined city bounding box, respectively. $grid\_size$ is the scale of discretization which is 15.625 m in our case. Because entropy rate metric requires binned data (binned to grid cells), we employed binned locations for convex hull and box-counting dimension as well. This has the disadvantage of reducing the resolution of measurements, but the advantage of reducing variability due to sensor noise.

3.2.5 Data Aggregation

iEpi and Ethica record sensor data for one minute in a duty cycle to save smartphone power. In the active one minute, more than one GPS record were recorded. The TAXI dataset updated current GPS position towards a server every 7 seconds. Following the steps of measuring mobility entropy rate in [69], we aggregated GPS data on the duty cycle for which data was collected. The first record in each duty cycle is taken as the representative location of that duty cycle. In Table 3.3, the record in the second row will be taken. There are other aggregation strategies, such as taking the mean or median of latitude and longitude as the location, or taking the first record with best accuracy.

3.2.6 Data Normalization

Normalization was carried out at several steps during data processing and analysis. The index of grids normalized within each dataset before extracting features with the maximum and minimum value for each dataset. The acquired features are normalized over all datasets for comparison with maximum and minimum value across all datasets. All normalizations follow the same strategy:

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{3.2}$$

where x is the current value, x_{max} and x_{min} are the maximum and minimum value of corresponding variable, respectively.

3.3 Detailed Configuration of Feature Calculations

Features were calculated over the duration of each study, for each participant. While smaller or sliding window timescales are possible, as the initial proof of concept, we restricted ourselves to the totality of data for each participant.

3.3.1 Activity Space Measures

Convex hull. Considering that specific locations encoded by grid index are required to calculate the convex hull of ten locations with longest dwell time, we also employed grid index for convex hull calculation for consistency. Because the grid index under base bin size (15.625 m) has been calculated in preprocessing, we directly applied the Python class ConvexHull from scipy.spatial which implements the Quickhull algorithm to the grid index of all records to get area of convex hull.

Convex hull of ten locations with longest dwell time. To extract ten locations with longest dwell time from GPS records, we used the index of each grid cell to aggregate locations. Side length of the grid is 250 meters instead of 15.625 meters in order to cluster movements within a place. Index of a grid of side length 250 m is 16 times of 15.625 m which makes it easy to transform from the index of grid of base size. A dwell starts when consecutive GPS records fall into the same grid, and ends when GPS location is outside the grid. Every dwell was summed to get the total dwell time in each grid cell. The 10 places with longest dwell time were extracted and the same ConvexHull class was employed to calculate the area of convex hull. Following the studies that often use size ranges between 200 meters and 500 meters [60, 90], we chose 250 meters as the maximum distance that a user can cover in a place to be considered as dwelling place.

Buffer area. As we are only interested in the area of buffered trip, we used the straight line between two locations as a simplification of real road network. Following [39], we took 200 meters as the buffer distance. UTM coordinates were used for calculating buffer area. Each two consecutive locations were connected with a straight line and treated as a segment of the whole path. Each segment was buffered using the buffer function in Python package shapely 1.6.4.post2. All buffers were combined using the cascaded_union function in the same package to get the overall buffer area.

3.3.2 Fractal Dimension

We employed Paul's implementation [68] of box-counting dimension which builds a n-Dimensional Tree [91]. Following this implementation, we used the grid index to aggregate GPS records to duty cycles as described in Preprocessing. The implementation can be rendered algorithmically as:

1. Remove repeated locations to avoid endless loops in the process of building n-Dimensional Tree.
2. Employ n-Dimensional Tree to produce a set of tuples of the form $\left[\log \frac{1}{\epsilon}, \log N(\epsilon)\right]$.
3. Fit a least-square regression line for $\log N(\epsilon)$ versus $\log \frac{1}{\epsilon}$.
4. Estimate box-counting dimension as the slope of regression line.

3.3.3 Entropy Rate

The constant terms in multiscale entropy rate model are latent variables, which are sum of different forms of apparent velocity and dwell time in each cell. Because GPS data are not collected continuously all the

time, but collected only one minute in every five minutes (if the duty cycle is 5 minutes), we don't have the complete location records. With the incomplete GPS records, we are not able to know the exact movement in each cell. For example, for a sequence represented as "AAA", an individual can move slowly across a cell or move fast and stop for a while in this cell to generate the sequence. So the terms can only be fitted instead of calculating for each cell directly using GPS data.

We followed the technique proposed in [70] to fit the constant terms C_1 to C_5. First, we calculated LZ-derived entropy rate lzH over pairs of (T, d) of the sequence of GPS locations represented as (x, y) according to Equation 2.10. The down-sampling intervals T of different datasets are shown in Table 3.6. The range of spatial quantization d is [15.625 m, 31.25 m, 62.5 m, 125 m, 250 m, 500 m, 1 km, 2 km, 4 km] following [70]. The resulted set of (T, d, L, lzH) were put into Eureqa [23] for data regression of Equation 2.11 to get the constant terms C_1 to C_5 for each participant.

Table 3.6: Down-sampling intervals of different datasets

Dataset	Down-sampling intervals
FSD	8 min, 40 min, 80 min, 2 hr, 4 hr, 8 hr
S9	5 min, 10 min, 30 min, 1 hr, 2 hr, 4 hr, 8 hr
S10	5 min, 10 min, 30 min, 1 hr, 2 hr, 4 hr, 8 hr
VIC	5 min, 10 min, 30 min, 1 hr, 2 hr, 4 hr, 8 hr
VAN	5 min, 10 min, 30 min, 1 hr, 2 hr, 4 hr, 8 hr
TAXI	1 min, 5 min, 10 min, 30 min, 1 hr, 2 hr, 4 hr, 8 hr

3.4 Classification

We extracted nine features in total from GPS locations of each participant. These features are area of convex hull (CH), area of convex hull of ten locations with longest dwell time (CH10), buffer area (BA), five constant terms ($C1$ to $C5$) of entropy rate considering varying spatial and temporal resolution, and box-counting dimension (DIM). To examine if the feature set could discriminate different populations, we trained a multi-class Support Vector Machine (SVM) model. We validated the effect of each feature on the overall classification performance by adding each feature incrementally to the classifier and rerunning the classification to compare their performance.

The models and feature selection were implemented with Python package scikit-learn 0.20.1. Because we hope to know the order of importance of each feature on the classifier, we employed the selectKBest function to get the complete order of importance according to ANOVA F-value metric. Other rankings of features are available for function selectKBest, such as principle component analysis (PCA), chi-squared, and mutual

information based rankings. Score function "chi2" computes chi-squared stats between each non-negative feature. Function "mutual_info_classif" estimates mutual information for a discrete target variable, and PCA based feature selection uses PCA function. The resulted ranking of different score functions are shown in Table 3.7. Features are listed in descending order of importance.

Table 3.7: Ranking of features based on different score functions

Score function	Ranking of features
f_classify	BA, C5, DIM, CH, C4, C2, C1, C3, CH10
CH2	C5, BA, C3, CH, DIM, C2, CH10, C4, C1
mutual_information	C5, CH, C3, C1, BA, DIM, C2, C4, CH10
PCA	C5, C4, CH10, CH, C3, C1, DIM, C2, BA

Although score functions generated different order of feature importance, $C5$ always shows its importance among all feature under all selection strategies. We used the order generated by selectKBest with score function "f_classif" for SVM model.

In the training step of SVM classifier, we tested the classifier on unbalanced datasets as well as balanced dataset with weights: [FSD: 28, S9: 5, S10: 7, VIC: 4, VAN: 10, TAXI: 1]. The weight is inversely proportional to the number of participants in that dataset. We used 75% of the dataset as training set and the remaining 25% as test set. We tested different kernels for SVM including linear, polynomial, RBF, and sigmoid kernels. The linear kernel outperformed other kernels and was used in the classification model. Finally, we employed 5-fold cross validation to derive the mean accuracy score and standard deviation value for the model performance evaluation.

3.5 Experiment Environment

All experiments were run on a MacBook Pro with 2.4 GHz Intel Core i5 and 8GB 1600 MHz DDR3 RAM. The general processing of GPS records and activity space features calculation were done using Python 3.6.7 with PyCharm 2017.1.3. Packages include Pandas 0.23.0, Numpy 1.15.4, scikit-learn 0.20.1, scipy 1.1.0, pyproj 1.9.5.1, folium 0.7.0, geopandas 0.4.0, shapely 1.6.4.post2, and matplotlib 3.0.2. Entropy rate and box-counting dimension were calculated using Paul's implementation [67, 68]. R packages ggplot2 3.1.0 and ggpubr 0.2 under R 3.5.0 were also used to draw boxplots in results. Functions aov and TukeyHSD in R were employed for one-way ANOVA test and Tukey's HSD test, respectively.

4 Results

To determine the utility of proposed features, ability of each feature to discriminate datasets in isolation was evaluated. By examining each feature in isolation, we can form hypotheses as to what phenomena it is sensitive to and selective for, and test those against our intuitions. To determine the utility of features as a feature set, a SVM model was trained to assign participants to the datasets from which they originated.

4.1 Single Feature Analysis

To demonstrate the discriminatory capability of each feature, we normalized each feature over all datasets and plotted the distribution of each metric in Figure 4.1. Each panel in the figure represents the distribution of each measure across full records for each participant in that study; that is each participant's value for each feature are calculated. Substantial differences between the relative values for each dataset are evident across the features, but no consistent pattern is evident, providing us with confidence that different features are enhancing different phenomenons in the observed spatial traces.

A two-step statistical analysis was applied to each metric. First, we used a one-way analysis of variance (ANOVA) to test the difference of distributions over all datasets, and subsequently, a Tukey's HSD test to discover the significantly different pairs of datasets. Because of the number of samples and discriminatory power of features, highly significant results from ANOVA is not meaningful for our analysis. We only report the results of Tukey's HSD tests in subsequent reporting, as summarized in Table 4.1. Data is reported to three decimal places for readability. Significance values close to 1 are reported as >0.999 and small significance values are reported as <0.001. Taking a standard 0.05 level as significant, all significant results are rendered in bold. Several trends are clearly evident from the table. FSD and VAN are always different, separated by both geography and demographics. All other datasets are almost always different from TAXI, except for VAN, as expected as taxi patterns would be different from individuals in most circumstances. Other differences in datasets are discussed in the following subsections.

4.1.1 Convex Hull and Its Variation

The distribution of area of convex hull across participants by dataset is shown in Figure 4.1-A. Tukey's HSD test shows that there is no significant difference between FSD, S9, S10, which are all collected in the same city (Saskatoon), as expected. There is no significant difference between datasets TAXI and VAN. The three Saskatoon datasets (FSD, S9, and S10) are different from TAXI and VAN. FSD is not different from VIC,

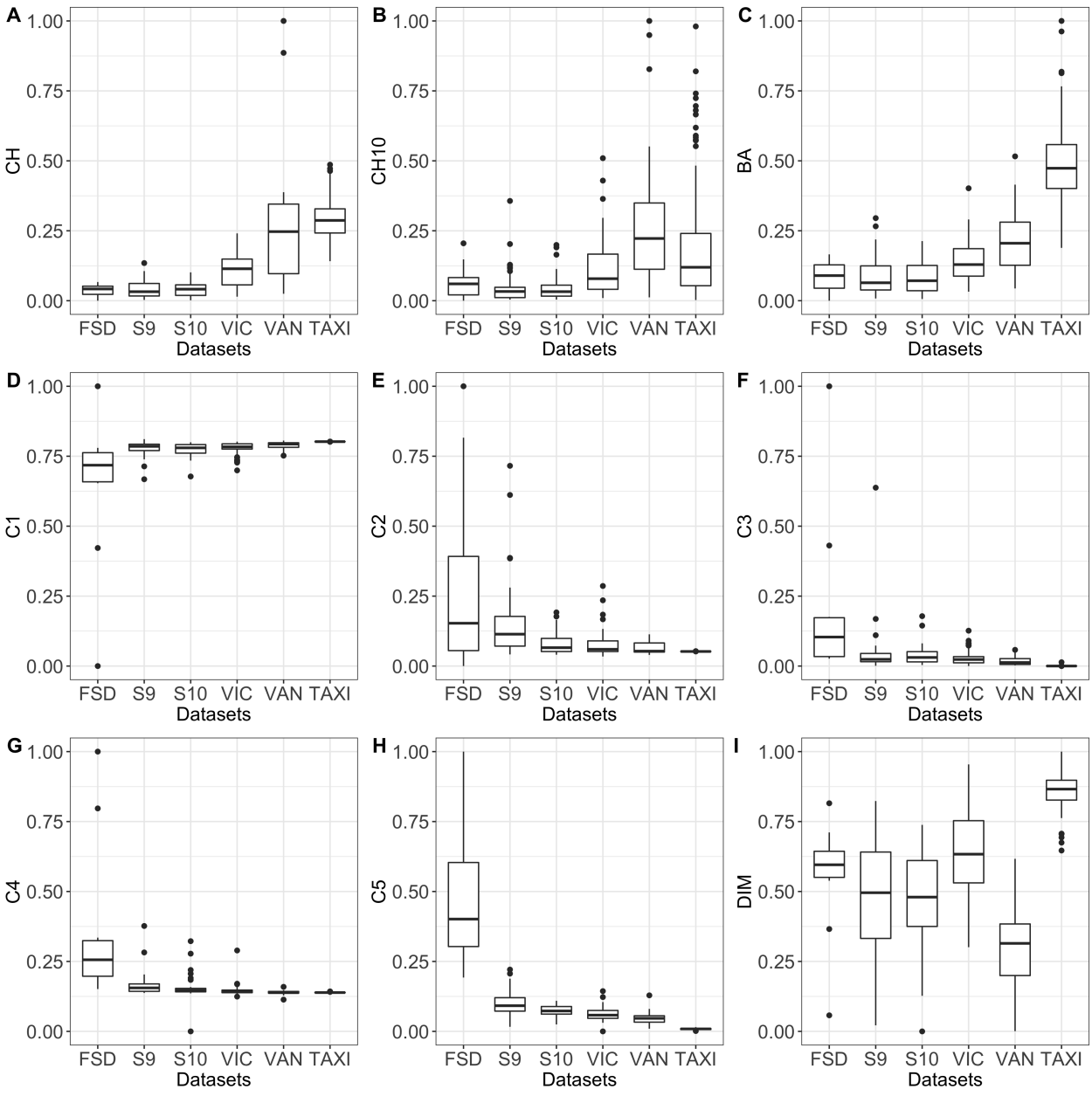

Figure 4.1: Distribution of each feature over all datasets

Table 4.1: P-value of Tukey's HSD test on all features

	Activity Space			**Entropy Rate**					**Fractal**
Dataset pairs	CH	CH10	BA	C1	C2	C3	C4	C5	DIM
FSD-S9	>0.999	0.997	>0.999	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	0.446
FSD-S10	>0.999	0.998	>0.999	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	0.289
FSD-VIC	0.074	0.946	0.580	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	0.286
FSD-VAN	**<0.001**	**<0.001**	**0.013**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**
FSD-TAXI	**<0.001**	0.240	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**
S9-S10	>0.999	>0.999	>0.999	0.994	**<0.001**	>0.999	0.955	**0.016**	0.993
S9-VIC	**<0.001**	0.084	**0.037**	>0.999	**<0.001**	0.659	0.209	**<0.001**	**<0.001**
S9-VAN	**<0.001**	**<0.001**	**<0.001**	0.895	**<0.001**	0.370	0.170	**<0.001**	**<0.001**
S9-TAXI	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	0.002	**<0.001**	**<0.001**
S10-VIC	**<0.001**	0.201	0.054	0.987	>0.999	0.878	0.870	0.809	**<0.001**
S10-VAN	**<0.001**	**<0.001**	**<0.001**	0.701	0.904	0.579	0.673	0.078	**<0.001**
S10-TAXI	**<0.001**	**<0.001**	**<0.001**	**<0.001**	0.109	**<0.001**	0.224	**<0.001**	**<0.001**
VIC-VAN	**<0.001**	**<0.001**	**0.036**	0.914	0.958	0.964	0.988	0.425	**<0.001**
VIC-TAXI	**<0.001**	**0.031**	**<0.001**	**<0.001**	0.067	**<0.001**	0.853	**<0.001**	**<0.001**
VAN-TAXI	0.401	**<0.001**	**<0.001**	0.465	0.949	0.462	>0.999	**<0.001**	**<0.001**

but S9 and S10, both studies in Saskatoon are different from VIC. Plausible explanations for these differences include that FSD participants were low income and had their behaviour constrained by mobility challenges due to income, while the VIC samples had their behaviour constrained by the topography of a coastal city [65]. S9 and S10 datasets differed from VIC because students may be less constrained in their activity space than low income individuals, perhaps partly due to the free bus pass program at the university [84].

The plot in Figure 4.1-B shows the distribution of area of convex hull built with ten locations with longest dwell time. Although this feature is also an area of convex hull, the post hoc results are different with the standard convex hull. VAN is different from all the other datasets, which is not surprising given its scale. TAXI is different from S9, S10, VIC, and VAN, which again is as expected as a taxi's top ten locations will be dictated by the whims of its customers more than the established geographic patterns of individuals. There is no significant differences between all Saskatoon datasets and VIC which is different from the results of standard convex hull. It shows that a tighter activity space between similar sized cities may be expected to be the similar. It is somewhat surprising that TAXI and FSD are not different. There is no significant

differences between any datasets collected in Saskatoon, the same result obtained from the standard convex hull.

4.1.2 Buffer Area

Buffer area reveals the space surrounding the locations visited by individuals during their daily movements, and is distinct from convex hull which instead describes the space circumscribed by the extent of people's movements. Figure 4.1-C is the distribution of buffer area normalized over all datasets. Again, significant difference were discovered between datasets [$F(5,487) = 305.1$, $p < 2e\text{-}16$]. According to the post hoc analysis, there is no significant difference between the Saskatoon datasets. VIC is also not significantly different from FSD. All the other pairs are significantly different.

4.1.3 Entropy Rate

The dependence of mobility entropy rate on spatial and temporal resolution $H(d, T)$ can be expressed by the five constant terms from [70]. We used R-squared value as measurement of how well a model fits the data. The distribution of R-squared value of each dataset is shown in Figure 4.2. According the the boxplot in Figure 4.2, GPS traces data of individuals from five datasets are well fitted with a mean R-squared value over 0.875. The high R-squared value indicates that the model proposed by Paul *et al.* [70] could match most datasets. R-squared values of fittings on taxicabs are relatively low compared to the other datasets. This may due to the less dense data points of a single taxicab compared to other datasets although the overall data quality is good. Note that R-squared value is not an intrinsically valid metric for non-linear fits. However, absent a better measure, we employed R-squared to determine relative fit quality.

Surfaces denoting the model with fitted constant terms, and the average LZ-derived entropy rate over each dataset, are shown in Figure 4.3. In Figure 4.3, d is in meters, T is in 10^4 seconds, and H is in bits. The wire frame surface is constructed with the average of entropy rate H calculated following Equation 2.11 with fitted constant terms. The scatter points are the average of LZ-derived entropy rate following Equation 2.10 over each dataset. Taxicabs exhibited the greatest entropy and FSD produced the lowest entropy. It is clear that the model proposed in [69] provides an accurate description of how mobility entropy rate varies across a wide range of spatial and temporal scales. As proposed in [70], these constants themselves can potentially serve as features. This is the first work we are aware of which examines the utility of these terms as features.

The distributions of $C1$-$C5$ are plotted in Figure 4.1-(D-H). ANOVA test reveals a significant difference in all constant terms among datasets. In Table 4.1, the following patterns are kept for all constant terms:

1. FSD is always significantly different from all the other datasets.

2. There is not a significant difference between VAN and S10, VIC and S10, and VAN and VIC.

3. In general, the constant term C5 is able to distinguish most pairs of datasets (12/15).

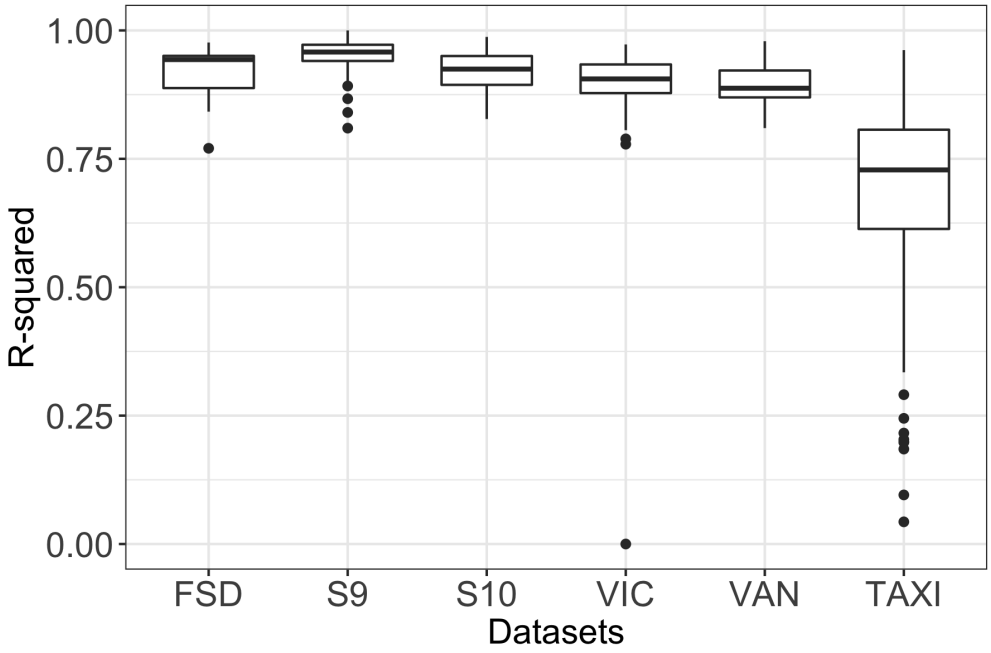

Figure 4.2: R-squared value of the fitting process for Equation 2.11

4.1.4 Fractal Dimension

Significant difference [F(5,487) = 250.1.5, p <2e-16] was found in fractal dimension among datasets. Tukey's HSD test shows that there is a significant difference between most pairs of datasets, but no significant difference between FSD and VIC, FSD and S9, FSD and S10, and S9 and S10. Datasets from similarly sized cities appear to have similar fractal dimension. This could be in part because a greater area provides more opportunity for more complex paths.

4.2 Application to Machine Learning

Mean accuracy of cross validation results of SVM with balanced data is 4% better than the model trained with unbalanced dataset. Following the feature selection step using selectKBest function in scikit-learn package, features ordered by importance are [BA, C5, DIM, CH, C4, C2, C1, C3, CH10]. We note that this is a greedy ordering of features, and other tests could generate other feature orderings. We built a sequence of SVM classifiers, increasing the number of included features, with the goal of assigning participants to datasets. After training the SVM, average accuracy was calculated from five-fold cross-validation. The accuracy score of five-fold cross validation is 0.36 (+/- 0.22) for a single feature (BA). When the second important feature (C5) was added to the feature set, the accuracy score increased to 0.72 (+/- 0.05). As expected, the accuracy increased monotonically with additional features, but with declining return. For the data examined here, most of the accuracy is gained after the first six to seven features were included, incorporating all three major

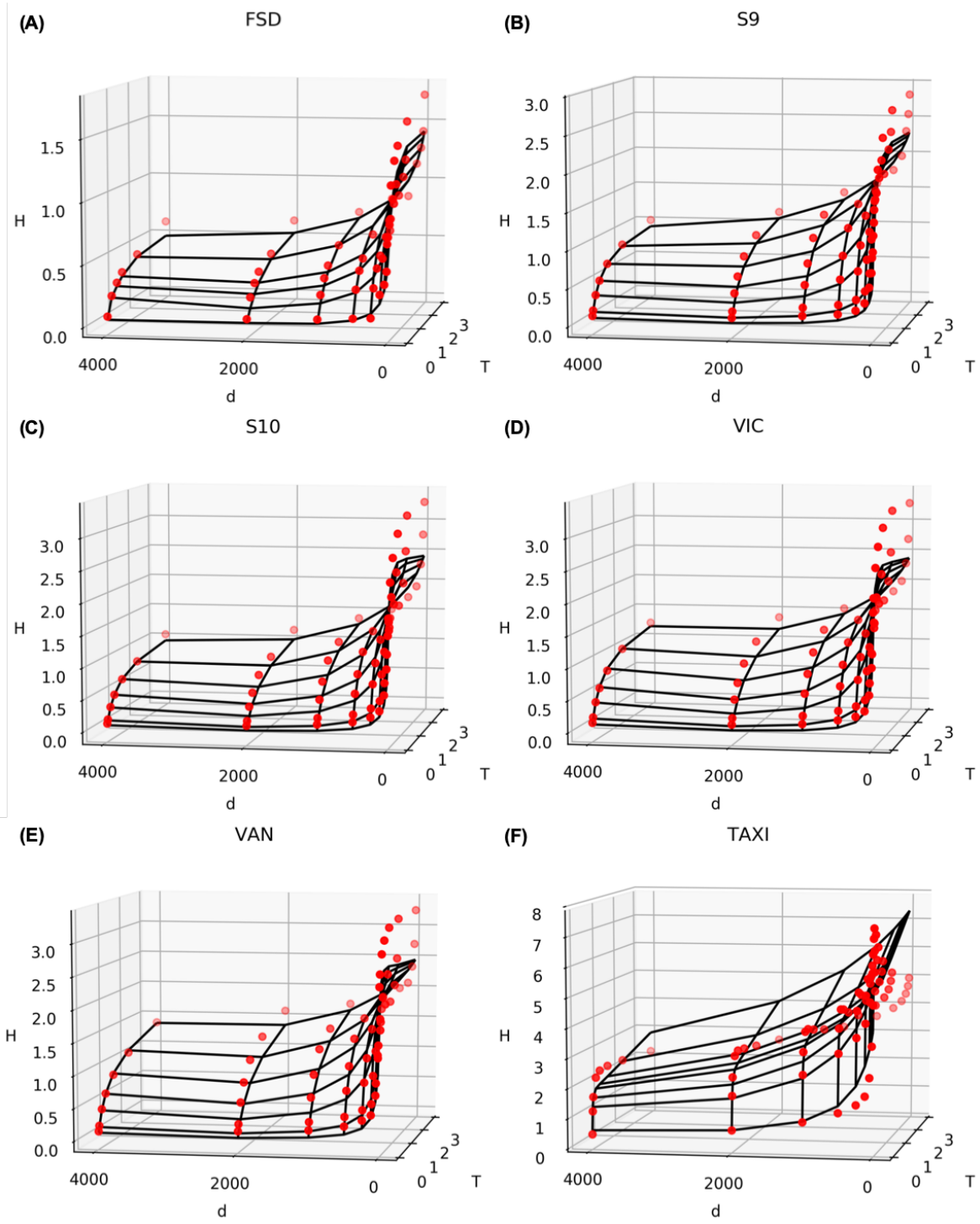

Figure 4.3: Entropy surface and empirical points of all datasets

classifications of features.

Confusion matrices of all SVM classifiers are shown in Figure 4.4. Increasing number of features were used in training SVM model. Above each confusion matrix, K denotes the count of features used in SVM, mean represents the mean accuracy of five-fold cross validation on training set, and std is the standard deviation of cross validation on training set.

It is worth noting, that unlike the statistical tests which endeavored to differentiate populations, SVM differentiated individuals. The model is attempting to answer the question: given an individual, which dataset is he from? This is a much more complex task. For a single feature, the classifier assigns most individuals to FSD and TAXI. The subsequent two features (C5, DIM), allow the near perfect assignment of participants to FSD, VAN and TAXI, with a reasonable number of true positives for VIC, but a large number of false positives, predominately arising from S9 and S10. Adding CH, C4 and C2 help to distinguish S9 from VIC, but S10 remains poorly classified, contributing a number of false positives to both S9 and VIC. C3 could distinguish some participants of S10 from S9 and VIC, but still in a poor condition. Additional features have marginal impact on the overall accuracy, and simply shift the false positives from S9 and S10 between each other and VIC.

4.3 Relationship between Food Preference and Spatial behaviour

Some of the datasets include surveys designed for different experiment problems, which can be exploited in this study. For example, the surveys in SHED10 posed questions about food purchasing behaviour and provided us an opportunity to explore how these features characterize groups within a study with respect to a specific human spatial behaviour.

In SHED10, participants were asked to report their food purchase during the study. 178 fast food purchases, 107 groceries for preparation later, 133 restaurant purchases and 33 pre-prepared food for consumption are reported in total. We focus on the grocery and fast food purchase because these two types indicate different life styles and health outcomes. There are 45 grocery purchases and 49 fast food purchases in total for 27 valid participants. We divide these participants into three groups according to their preferred food type, *i.e.*, more grocery, more fast food and no preference. Finally, 12 participants, 13 participants, and 2 participants are assigned as preferring fast food, preferring grocery and no preference, respectively. Because the no preference group had only 2 participants, we only analyzed the difference of nine features between preferring fast food group and preferring grocery group.

The distribution of all features are shown in Figure 4.5 and the results of one-way ANOVA are reported in Table 4.2. Because multiple tests are performed simultaneously on the same datasets, we need to perform corrections to p-value from ANOVA, such as Bonferroni correction. The Bonferroni correction compensates the increase of Type I error in multiple tests by testing each hypothesis at a significance level of $\frac{\alpha}{m}$ where α is the desired overall alpha level, such as 0.05, and m is the number of hypotheses, which is 9 in our case.

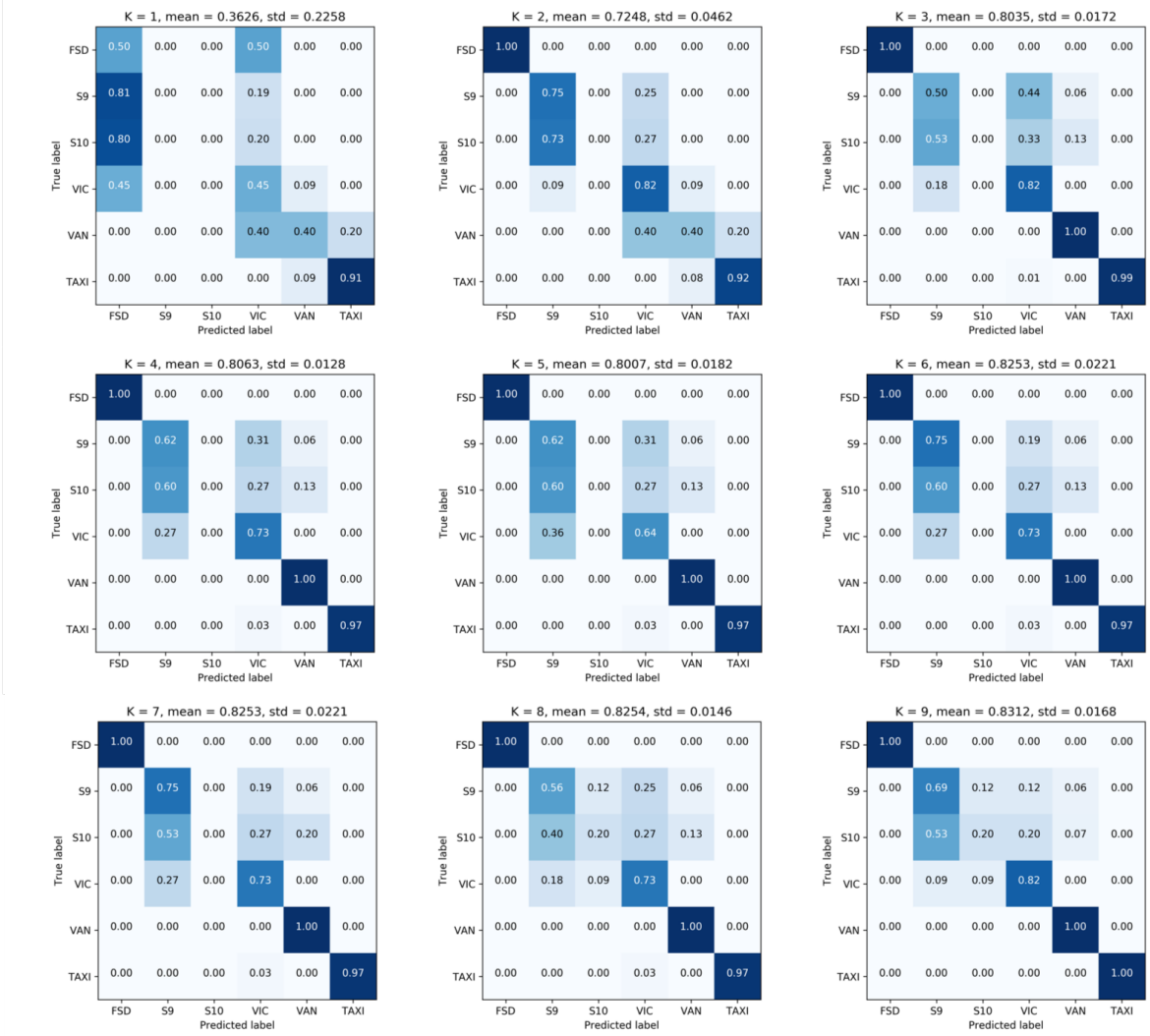

Figure 4.4: Confusion matrix of SVM classifier on test set

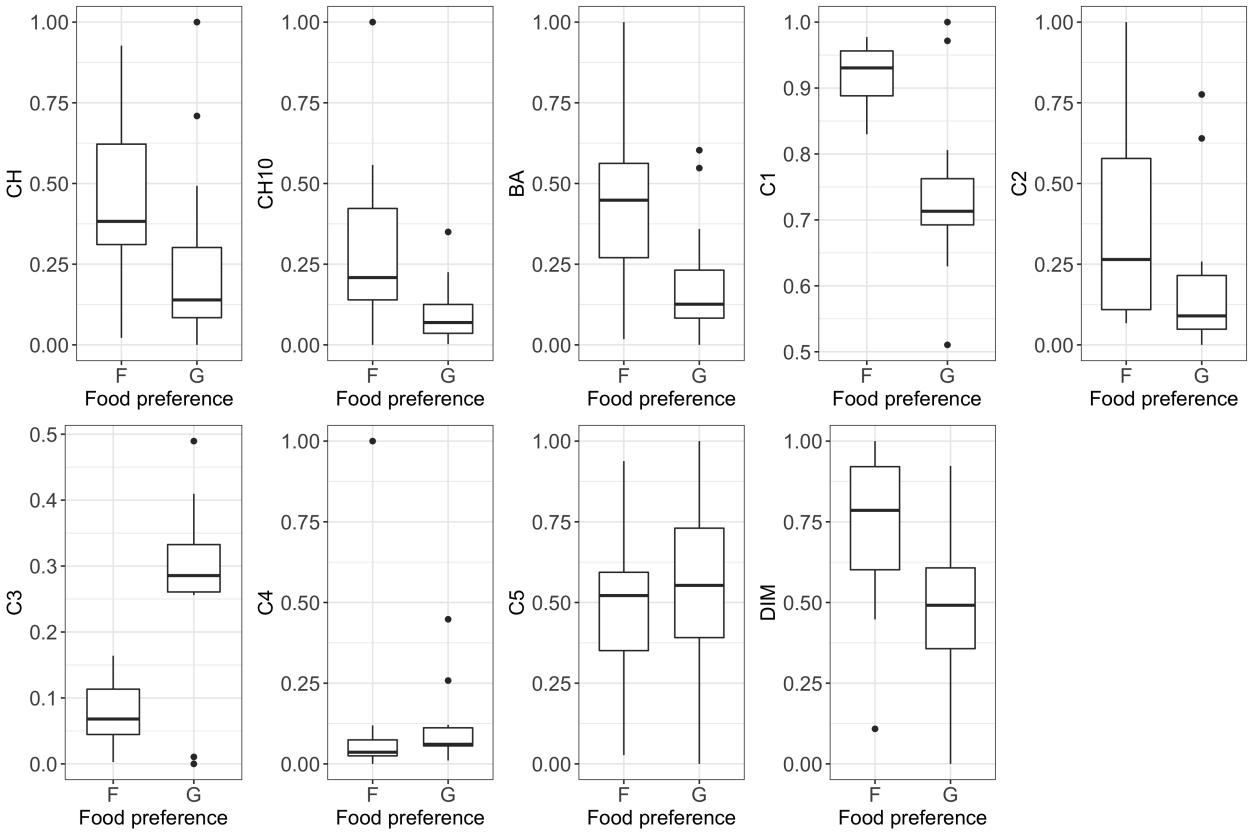

Figure 4.5: Feature distribution of participants with different food preference

So p-value needs to be at a significance level of 0.0056, and the two food preference groups are significantly different on features C1 and C3.

Table 4.2: One-way ANOVA output on all features

	CH	CH10	BA	C1	C2	C3	C4	C5	DIM
Sum of Squares	0.2018	0.2636	0.3345	0.1873	0.2102	0.2131	0.0008	0.0094	0.3463
Mean Squares	0.20175	0.26356	0.3345	0.18735	0.21018	0.2131	0.00085	0.00936	0.3463
F-values	2.181	5.872	5.83	18.3	2.358	18.86	0.018	0.122	4.763
P-value	0.154	0.0241	0.0245	0.0003	0.139	0.00026	0.893	0.73	0.04

5 Discussion

5.1 Summary

Researchers from different disciplines have proposed a variety of methods to quantify and qualify spatial behaviour. However, there isn't a standard feature set which can provide a comprehensive description of spatial behaviour and could be used for classification or clustering. In this thesis, we constructed a potential feature set composed of convex hull and its variation, buffer area, fractal dimension and constant terms of multiscale entropy rate. We extracted the feature set of six datasets collected from varying demographic and regions, and examined how these features describe datasets, how they differentiate each dataset, and how they relate to other individual characteristic such as food preference.

The results reveal the ability of proposed feature set to differentiate different datasets. As the results of the SVM model show, no single feature could distinguish all datasets successfully. Although results of Tukey's HSD imply that DIM and C5 could distinguish most pairs of datasets, they still fail to separate some pairs. For example, C5 could not separate S10 and VIC, and DIM fails to separate FSD and S9, FSD and S10, FSD and VIC. In general, the fact of any single feature couldn't distinguish all datasets is reasonable. Because a single feature can only characterize a specific aspect of spatial behaviour, while different datasets are not always differ on the same aspect. For example, datasets S9 and S10 have same demographic and geographic background, so they are expected to have similar spatial range which is quantified by convex hull and buffer area. But the difference between these two datasets, i.e., seasons (Fall and Winter) is possibly revealed in the dwell-time proportional constants in the scale free entropy rate calculation. According to the results and each feature's property, measures quantifying activity space were indicative for the city where a dataset was collected, while fractal dimension emphasizes demographic differences between datasets, i.e., taxicabs, cyclists and low income people experience more fractal movement patterns than datasets of university students.

FSD, S9, and S10 show difference on CH and CH10. With CH, these three datasets from same city show quite similar distribution, while the value of CH10 of FSD turns to be larger than that of S9 and S10. We can interpret this difference as undergraduate students' mostly frequently visited places are within a smaller area than FSD although they occasionally visit some faraway places. The difference between BA and convex hull can be seen on VAN and TAXI. Because Vancouver is a larger city than Rome, the maximum value of CH and CH10 in VAN is larger than that of TAXI, which reveals the area of cities. This pattern changes on BA, because BA mainly measures the overall trip length, and taxicabs have longer travel distance than individuals as expected. BA is not always proportional to trip length if trip segments overlap because the

buffer of overlapped segments will be merged.

Compared to activity space related features and fractal dimension which have explicit meanings, scale-free entropy rate measures which correspond to five constant terms C1 to C5 are more difficult to interpret. We could understand the entropy rate results by mapping the constant terms to the value being summed over cells. According to the derivation in Equation 2.11, $C1$ is one over the squared velocity, $C2$ is the squared dwell time, $C3$ is the quotient of dwell time and velocity, $C4$ is one over the velocity, and $C5$ is dwell time. As shown in Table 4.1, FSD is different from all the other datasets across all constants. This may reveal the income difference between FSD and the others, it may also caused by the violation of Paul's assumption for Equation 2.11. In [70], the traces are assumed to be sufficiently dense and long for the Lz-derived entropy approximation to converge, and the fits of Equation 2.11 to be meaningful. Considering the largest duty cycle (8 mins VS. 5 mins) and the poorer data quality of FSD compared to the other dataset, the significant difference in Tukey's results may due to the violation of one or more of Paul's assumption. Among all five constant terms, $C5$, which is the sum of dwell time, is the most powerful feature and is able to differentiate most pairs of datasets. It could differentiate TAXI dataset from all the other datasets which is reasonable because taxicabs would have significantly different dwelling patterns than individuals. $C5$ could also differentiates S9 from S10 which are much more similar than the other pairs. This may reveal the different dwell patterns for individuals in fall and winter in Saskatoon where the temperature in winter is quite low. $C2$ could also distinguish S9 from S10 for the same reason. $C1$ and $C3$ have the same power on differentiating datasets.

The two-step statistical analysis reveals the ability of each single feature to distinguish different datasets, and building a SVM model with these features shows the discrimination ability of the features as a feature set. The feature sorting is done with f_classify function in scikit-learn package which uses ANOVA F-value between label and feature. Different stratifications could generate different importance order of features, and could be employed for other datasets. Based on the results of feature selection, a series of SVM model were built to determine the extent to which each feature provided additional discriminatory power between datasets. A maximum accuracy score of 0.85 on test set was achieved after all nine features were added. The increasing accuracy of SVM model with increasing number of features indicates that these features encode different characteristics of spatial behaviour, and is powerful in discrimination as a whole set. Confusion matrix analysis showed clearly the performance of features on distinguish different datasets. Overall, SVM model with the first three features, i.e., BA, $C5$, and DIM assigned participants from FSD, VAN and TAXI with excellent accuracy. Mean accuracy of 0.80 of cross validation and the high accuracy on FSD, TAXI, and VAN indicate that the first three features covering all three disciplines are able to characterize these three datasets relatively comprehensively. Those same features were not able to distinguish more similar datasets S9, S10, and VIC. While VIC had a reasonable true positive rate, it was confounded by false positives from S9 and S10. After adding the first three features, the increase of accuracy slows down. Confusion matrices show that newly added features are attempting to resolve hard cases among S9, S10, and VIC. Even with all nine

features, more than half of participants from S10 are assigned to S9. The main finding from the classification model was that while the features could distinguish populations, it is only reliable to differentiate individuals from highly discriminative populations. Besides, the quite small size of FSD may cause incorrect results of SVM model.

5.2 Contributions

The study described in this thesis is an attempt towards a standard feature set for characterizing spatial behaviour. It has the following contributions to the community:

Establish a feature set and verify its utility in distinguish different populations. According to the analysis in this study, we could demonstrate the advantages of leveraging a standardized feature set. First, the standard feature set enables the generalization of findings across both existing studies and newly published datasets, such as studies from different populations. The features described here could be employed not only in human's spatial movement data, but a wide variety of spatial data with sufficiently dense location-derived streams.

Provide a benchmark for spatial behaviour feature designing. The features provide a standard baseline of feature discrimination. An additional spatial feature can be considered useful only if it can provide a more efficient or distinct representation of spatial behaviour than the features already included in the feature set.

Build a package for feature extraction from GPS datasets. With the Python code built in the study, researchers are freed from fraught feature engineering and can focus on the results analysis. Absent strong hypothesis about spatial behaviour inherent in the data, this quick analysis could provide initial insight to further data exploration.

First to employ fractal dimension for human spatial behaviour. Convex hull and buffer area are widely used and comprehensible features in GIScience. Mobility entropy rate is more novel than convex hull, but has a growing body of literature and straightforward applications on measuring human mobility. As far as we know, we are the first one to apply fractal dimension to characterize human spatial behaviour in our basic work [98] for this thesis. The statistical analysis and classification model both reveal the ability of fractal dimension in differentiating populations.

First to employ scale free entropy rate coefficients as features for human spatial behaviour. Entropy rate has been successfully employed to describe human spatial behaviour, but we're the first one to use the latest scale free entropy rate coefficients [70] as features. The scale free entropy rate enables comparisons between datasets with different spatial and temporal rates. The coefficients help researchers understand results by mapping the constants to the value being summed over cells.

5.3 Shortcomings

While the work presented here has made notable contributions to the characterization of human spatial behaviour, our work still have limitations to be addressed and substantial opportunities for future study.

Meaning of Features In this study, we have made an initial attempt to interpret the possible phenomenon that a single feature or combination of features represent. However, we are not able to make definitive conclusions about their meanings. In the future, a more comprehensive and accurate discussion about how and which features describe the different characteristic of spatial behaviour could be analyzed with controlled experiments.

Location independent All the features employed in this thesis do not involve any location information. We treat locations as points on a plane and treat trajectory as a curve or a string of location symbols. None of these methods include information about these locations and what people do at these places. The lack of location information may effect interpretation of features. This shortcoming can be overcome using surveys or other location information such as Google map.

Datasets Although we evaluated the validity of the feature set over six datasets from different geographic area and different demographics, the overall diversity of populations is still limited. Five datasets were collected in Canadian cities and TAXI is from Italy. All cities involved are from developed countries. Second, the features described in the study are not limited to human spatial movement data. Feature set can be applied to any GPS or location trace, including animal or complex physical paths as analyzed in Paul *et al.* [70], or even traces through the virtual worlds described in computer games, if assumptions about underlying data for deriving features are met.

Feature Selection As the first trying to build a feature set for spatial behaviour, we evaluated the relatively simple versions of each feature, which have known limitations, and can be addressed in the future work. The features we discussed in the study are only a small portion of the potentially available features. We are expecting novel and powerful features to be added to the feature set in the future. A growing library of validated, and phenomenally meaningful features would highly benefit the community.

Involvement of other researchers As the spatial behaviour is an important foundation in a lot of research problems from different disciplines, and the purpose of building a standard feature set is to facilitate all related research, it will be beneficial to involve user evaluation. Researchers in GIScience, Health, or Zoology could validate the effectiveness and generalization of the features in their domains.

6 Conclusion

In this thesis, we have described a standard feature set for spatial behaviour analysis based on GPS traces. The feature set includes nine features drawn from three distinct mathematical disciplines: geometry, information theory, and fractal analysis. Six datasets with different core demographic and/or geographic features were employed to validate the proposed feature set. These features were evaluated separately with two-step statistical analysis for their ability to distinguish datasets. The results indicates that while each single feature could discriminate a different subset of the six datasets, no single feature is able to distinguish every pair of datasets. Further, the classification task with SVM model reveals that combination of features as a whole feature set could assign individuals to the correct datasets with an accuracy of 0.85. The standard feature set will benefit the community by providing rapid exploration, standard analysis and reporting of data. It enables a faster, easier and reliable process of location data, and an interpretable, comparable and extendable understanding of spatial behaviour of human and other objects.

References

[1] Hervé Abdi and Lynne J Williams. Tukey's honestly significant difference (hsd) test. *Encyclopedia of Research Design. Thousand Oaks, CA: Sage*, pages 1–5, 2010.

[2] Raul Amici, Marco Bonola, Lorenzo Bracciale, Antonello Rabuffi, Pierpaolo Loreti, and Giuseppe Bianchi. Performance assessment of an epidemic protocol in vanet using real traces. *Procedia Computer Science*, 40:92–99, 2014.

[3] Theo A Arentze, Harmen Oppewal, and Harry JP Timmermans. A multipurpose shopping trip model to assess retail agglomeration effects. *Journal of Marketing Research*, 42(1):109–115, 2005.

[4] Daniel Ashbrook and Thad Starner. Using gps to learn significant locations and predict movement across multiple users. *Personal and Ubiquitous computing*, 7(5):275–286, 2003.

[5] Kay W Axhausen, Andrea Zimmermann, Stefan Schönfelder, Guido Rindsfüser, and Thomas Haupt. Observing the rhythms of daily life: A six-week travel diary. *Transportation*, 29(2):95–124, 2002.

[6] Trevor C Bailey and Anthony C Gatrell. *Interactive spatial data analysis*, volume 413. Longman Scientific & Technical Essex, 1995.

[7] C Bradford Barber, David P Dobkin, and Hannu Huhdanpaa. The quickhull algorithm for convex hulls. *ACM Transactions on Mathematical Software (TOMS)*, 22(4):469–483, 1996.

[8] Richard Becker, Ramón Cáceres, Karrie Hanson, Sibren Isaacman, Ji Meng Loh, Margaret Martonosi, James Rowland, Simon Urbanek, Alexander Varshavsky, and Chris Volinsky. Human mobility characterization from cellular network data. *Communications of the ACM*, 56(1):74–82, 2013.

[9] Ravi Bhoraskar, Nagamanoj Vankadhara, Bhaskaran Raman, and Purushottam Kulkarni. Wolverine: Traffic and road condition estimation using smartphone sensors. In *Communication Systems and Networks (COMSNETS), 2012 Fourth International Conference on*, pages 1–6. IEEE, 2012.

[10] James Biagioni, Tomas Gerlich, Timothy Merrifield, and Jakob Eriksson. Easytracker: automatic transit tracking, mapping, and arrival time prediction using smartphones. In *Proceedings of the 9th ACM Conference on Embedded Networked Sensor Systems*, pages 68–81. ACM, 2011.

[11] Lorenzo Bracciale, Marco Bonola, Pierpaolo Loreti, Giuseppe Bianchi, Raul Amici, and Antonello Rabuffi. CRAWDAD dataset roma/taxi (v. 2014-07-17). Downloaded from `https://crawdad.org/roma/taxi/20140717/taxicabs`, July 2014. traceset: taxicabs.

[12] Ron N Buliung and Pavlos S Kanaroglou. Urban form and household activity-travel behavior. *Growth and Change*, 37(2):172–199, 2006.

[13] Ronald N Buliung and Pavlos S Kanaroglou. A gis toolkit for exploring geographies of household activity/travel behavior. *Journal of Transport Geography*, 14(1):35–51, 2006.

[14] Francesco Camastra. Data dimensionality estimation methods: a survey. *Pattern recognition*, 36(12):2945–2954, 2003.

[15] Spencer Chainey, Lisa Tompson, and Sebastian Uhlig. The utility of hotspot mapping for predicting spatial patterns of crime. *Security journal*, 21(1-2):4–28, 2008.

[16] Basile Chaix, Yan Kestens, Camille Perchoux, Noëlla Karusisi, Juan Merlo, and Karima Labadi. An interactive mapping tool to assess individual mobility patterns in neighborhood studies. *American journal of preventive medicine*, 43(4):440–450, 2012.

[17] T Chambers, AL Pearson, I Kawachi, Z Rzotkiewicz, J Stanley, M Smith, C Ni Mhurchu, L Signal, et al. Kids in space: Measuring children's residential neighborhoods and other destinations using activity space gps and wearable camera data. *Social Science & Medicine*, 193:41–50, 2017.

[18] Donald R Chand and Sham S Kapur. An algorithm for convex polytopes. *Journal of the ACM (JACM)*, 17(1):78–86, 1970.

[19] John D Corbit and David J Garbary. Fractal dimension as a quantitative measure of complexity in plant development. *Proc. R. Soc. Lond. B*, 262(1363):1–6, 1995.

[20] DJ Coughlin, JR Strickler, and B Sanderson. Swimming and search behaviour in clownfish, amphiprion perideraion, larvae. *Animal Behaviour*, 44:427–440, 1992.

[21] Clodoveu A Davis Jr, Gisele L Pappa, Diogo Rennó Rocha de Oliveira, and Filipe de L. Arcanjo. Inferring the location of twitter messages based on user relationships. *Transactions in GIS*, 15(6):735–751, 2011.

[22] Jennifer Dill, Nathan McNeil, Joseph Broach, and Liang Ma. Bicycle boulevards and changes in physical activity and active transportation: Findings from a natural experiment. *Preventive medicine*, 69:S74–S78, 2014.

[23] Renáta Dubčáková. Eureqa: software review. *Genetic programming and evolvable machines*, 12(2):173–178, 2011.

[24] Nathan Eagle and Alex Sandy Pentland. Reality mining: sensing complex social systems. *Personal and ubiquitous computing*, 10(4):255–268, 2006.

[25] Yingling Fan and Asad J Khattak. Urban form, individual spatial footprints, and travel: Examination of space-use behavior. *Transportation Research Record*, 2082(1):98–106, 2008.

[26] Ronald A Fisher. Xv.—the correlation between relatives on the supposition of mendelian inheritance. *Earth and Environmental Science Transactions of the Royal Society of Edinburgh*, 52(2):399–433, 1919.

[27] Fabien Girardin, Francesco Calabrese, Filippo Dal Fiore, Carlo Ratti, and Josep Blat. Digital footprinting: Uncovering tourists with user-generated content. *IEEE Pervasive computing*, 7(4):36–43, 2008.

[28] Reginald G Golledge. *Wayfinding behavior: Cognitive mapping and other spatial processes*. JHU press, 1999.

[29] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *nature*, 453(7196):779, 2008.

[30] Ronald L. Graham. An efficient algorithm for determining the convex hull of a finite planar set. *Info. Pro. Lett.*, 1:132–133, 1972.

[31] Peter Grassberger and Itamar Procaccia. Measuring the strangeness of strange attractors. *Physica D: Nonlinear Phenomena*, 9(1-2):189–208, 1983.

[32] Tony H Grubesic. On the application of fuzzy clustering for crime hot spot detection. *Journal of Quantitative Criminology*, 22(1):77, 2006.

[33] Torsten Hägerstraand. What about people in regional science? *Papers in regional science*, 24(1):7–24, 1970.

[34] Susan Hanson and O James Huff. Systematic variability in repetitious travel. *Transportation*, 15(1-2):111–135, 1988.

[35] Gabriella M Harari, Nicholas D Lane, Rui Wang, Benjamin S Crosier, Andrew T Campbell, and Samuel D Gosling. Using smartphones to collect behavioral data in psychological science: opportunities, practical considerations, and challenges. *Perspectives on Psychological Science*, 11(6):838–854, 2016.

[36] Mohammad Hashemian, Dylan Knowles, Jonathan Calver, Weicheng Qian, Michael C Bullock, Scott Bell, Regan L Mandryk, Nathaniel Osgood, and Kevin G Stanley. iepi: an end to end solution for collecting, conditioning and utilizing epidemiologically relevant data. In *Proceedings of the 2nd ACM international workshop on Pervasive Wireless Healthcare*, pages 3–8. ACM, 2012.

[37] Felix Hausdorff. Dimension und äußeres maß. *Mathematische Annalen*, 79(1-2):157–179, 1918.

[38] Juan C Herrera, Daniel B Work, Ryan Herring, Xuegang Jeff Ban, Quinn Jacobson, and Alexandre M Bayen. Evaluation of traffic data obtained via gps-enabled mobile phones: The mobile century field experiment. *Transportation Research Part C: Emerging Technologies*, 18(4):568–583, 2010.

[39] Jana A Hirsch, Meghan Winters, Philippa Clarke, and Heather McKay. Generating gps activity spaces that shed light upon the mobility habits of older adults: a descriptive analysis. *International journal of health geographics*, 13(1):51, 2014.

[40] Frank E Horton and David R Reynolds. Action space formation: a behavioral approach to predicting urban travel behavior. *Highway Research Record*, (322), 1970.

[41] Frank E Horton and David R Reynolds. Effects of urban spatial structure on individual behavior. *Economic Geography*, 47(1):36–48, 1971.

[42] Ethica Data Services Inc. Ethica data. `https://www.ethicadata.com/`, 2018.

[43] Sibren Isaacman, Richard Becker, Ramón Cáceres, Stephen Kobourov, Margaret Martonosi, James Rowland, and Alexander Varshavsky. Ranges of human mobility in los angeles and new york. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2011 IEEE International Conference on*, pages 88–93. IEEE, 2011.

[44] Raymond Austin Jarvis and Edward A Patrick. Clustering using a similarity measure based on shared near neighbors. *IEEE Transactions on computers*, 100(11):1025–1034, 1973.

[45] Seungwoo Kang, Jinwon Lee, Hyukjae Jang, Youngki Lee, Souneil Park, and Junehwa Song. A scalable and energy-efficient context monitoring framework for mobile personal sensor networks. *IEEE Transactions on Mobile Computing*, 9(5):686–702, 2010.

[46] Balázs Kégl. Intrinsic dimension estimation using packing numbers. In *Advances in neural information processing systems*, pages 697–704, 2003.

[47] Yan Kestens, Meghan Winters, Daniel Fuller, Scott Bell, Janelle Berscheid, Ruben Brondeel, Michael Cantinotti, Geetanjali Datta, Lise Gauvin, Margot Gough, et al. Interact: A comprehensive approach to assess urban form interventions through natural experiments. *BMC public health*, 19(1):51, 2019.

[48] Dylan L Knowles, Kevin G Stanley, and Nathaniel D Osgood. A field-validated architecture for the collection of health-relevant behavioural data. In *2014 IEEE International Conference on Healthcare Informatics*, pages 79–88. IEEE, 2014.

[49] Mei-Po Kwan. Gender and individual access to urban opportunities: a study using space–time measures. *The Professional Geographer*, 51(2):210–227, 1999.

[50] Mei-Po Kwan. Interactive geovisualization of activity-travel patterns using three-dimensional geographical information systems: a methodological exploration with a large data set. *Transportation Research Part C: Emerging Technologies*, 8(1-6):185–203, 2000.

[51] Mei-Po Kwan and Xiao-Dong Hong. Network-based constraints-oriented choice set formation using gis. *Geographical Systems*, 5:139–162, 1998.

[52] Nicholas D Lane, Emiliano Miluzzo, Hong Lu, Daniel Peebles, Tanzeem Choudhury, and Andrew T Campbell. A survey of mobile phone sensing. *IEEE Communications magazine*, 48(9), 2010.

[53] Young-Seol Lee and Sung-Bae Cho. Activity recognition using hierarchical hidden markov models on a smartphone with 3d accelerometer. In *International Conference on Hybrid Artificial Intelligence Systems*, pages 460–467. Springer, 2011.

[54] Renaud Lopes and Nacim Betrouni. Fractal and multifractal analysis: a review. *Medical image analysis*, 13(4):634–649, 2009.

[55] Benoit B Mandelbrot. Les objets fractals: forme, hasard et dimension. 1975.

[56] Benoit B Mandelbrot. *The fractal geometry of nature*, volume 173. WH freeman New York, 1983.

[57] Benoit B Mandelbrot, Form Fractals, and WH Chance. San francisco, 1977.

[58] Liang Mao and Dawn Nekorchuk. Measuring spatial accessibility to healthcare for populations with multiple transportation modes. *Health & place*, 24:115–122, 2013.

[59] Liam McNamara, Cecilia Mascolo, and Licia Capra. Media sharing based on colocation prediction in urban transport. In *Proceedings of the 14th ACM international conference on Mobile computing and networking*, pages 58–69. ACM, 2008.

[60] Raul Montoliu and Daniel Gatica-Perez. Discovering human places of interest from multimodal mobile phone data. In *Proceedings of the 9th International Conference on Mobile and Ubiquitous Multimedia*, page 12. ACM, 2010.

[61] Mirco Nanni, Roberto Trasarti, Barbara Furletti, Lorenzo Gabrielli, Peter Van Der Mede, Joost De Bruijn, Erik De Romph, and Gerard Bruil. Transportation planning based on gsm traces: a case study on ivory coast. In *Citizen in Sensor Networks*, pages 15–25. Springer, 2014.

[62] Osagie Osemwegie and Kevin Stanley. Scalable and energy efficient software architecture for human behavioral measurement. In *Information Technology, Electronics and Mobile Communication Conference (IEMCON), 2016 IEEE 7th Annual*, pages 1–8. IEEE, 2016.

[63] Nathaniel D Osgood, Tuhin Paul, Kevin G Stanley, and Weicheng Qian. A theoretical basis for entropy-scaling effects in human mobility patterns. *PloS one*, 11(8):e0161630, 2016.

[64] Edward Ott. *Chaos in dynamical systems*. Cambridge university press, 2002.

[65] Yoo Park and Mei-Po Kwan. Multi-contextual segregation and environmental justice research: Toward fine-scale spatiotemporal approaches. *International journal of environmental research and public health*, 14(10):1205, 2017.

[66] Zachary Patterson and Steven Farber. Potential Path Areas and Activity Spaces in Application: A Review. *Transport Reviews*, 35(6):679–700, 2015.

[67] Tuhin Paul. Calculation of lempel-ziv entropy rate. `https://github.com/tuhinpaul/lz_entropy_rate`.

[68] Tuhin Paul. Mobility analysis. `https://git.cs.usask.ca/tuhin.paul/mobility-analysis.git`, 2017.

[69] Tuhin Paul et al. *Modeling Human Mobility Entropy as a Function of Spatial and Temporal Quantizations*. PhD thesis, University of Saskatchewan, 2017.

[70] Tuhin Paul, Kevin G Stanley, and Nathaniel D Osgood. Multiscale entropy rate analysis of complex mobile agents. *Royal Society open science*, 5(10):180488, 2018.

[71] Marie-Pier Pelletier, Martin Trépanier, and Catherine Morency. Smart card data use in public transit: A literature review. *Transportation Research Part C: Emerging Technologies*, 19(4):557–568, 2011.

[72] Weicheng Qian, Kevin G Stanley, and Nathaniel D Osgood. The impact of spatial resolution and representation on human mobility predictability. In *International Symposium on Web and Wireless Geographical Information Systems*, pages 25–40. Springer, 2013.

[73] Kiran K Rachuri, Mirco Musolesi, Cecilia Mascolo, Peter J Rentfrow, Chris Longworth, and Andrius Aucinas. Emotionsense: a mobile phones based adaptive platform for experimental social psychology research. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, pages 281–290. ACM, 2010.

[74] R. Rai, Michael Balmer, Marcel Rieser, V. Vaze, Stefan Schönfelder, and Kay Axhausen. Capturing Human Activity Spaces: New Geometries. *Transportation Research Record: Journal of the Transportation Research Board*, 2021(October 2015):70–80, 2007.

[75] Tye Rattenbury, Nathaniel Good, and Mor Naaman. Towards automatic extraction of event and place semantics from flickr tags. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 103–110. ACM, 2007.

[76] Jennifer Rogalsky. The working poor and what gis reveals about the possibilities of public transit. *Journal of Transport Geography*, 18(2):226–237, 2010.

[77] Stefan Schönfelder and Kay W Axhausen. Activity spaces: measures of social exclusion? *Transport policy*, 10(4):273–286, 2003.

[78] Stefan Schönfelder, Kay W Axhausen, Nicolas Antille, and Michel Bierlaire. Exploring the potentials of automatically collected gps data for travel behaviour analysis: A swedish data source. *Arbeitsberichte Verkehrs-und Raumplanung*, 124, 2002.

[79] Thomas Schürmann and Peter Grassberger. Entropy estimation of symbol sequences. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 6(3):414–427, 1996.

[80] Martine Shareck, Yan Kestens, and Lise Gauvin. Examining the spatial congruence between data obtained with a novel activity location questionnaire, continuous gps tracking, and prompted recall surveys. *International journal of health geographics*, 12(1):40, 2013.

[81] Jill E Sherman, John Spencer, John S Preisser, Wilbert M Gesler, and Thomas A Arcury. A suite of methods for representing activity space in a healthcare accessibility study. *International journal of health geographics*, 4(1):24, 2005.

[82] Gavin Smith, Romain Wieser, James Goulding, and Duncan Barrack. A refined limit on the predictability of human mobility. In *Pervasive Computing and Communications (PerCom), 2014 IEEE International Conference on*, pages 88–94. IEEE, 2014.

[83] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

[84] Kevin Stanley, Scott Bell, L Kurt Kreuger, Priyasree Bhowmik, Narjes Shojaati, Alexa Elliott, and Nathaniel D Osgood. Opportunistic natural experiments using digital telemetry: a transit disruption case study. *International Journal of Geographical Information Science*, 30(9):1853–1872, 2016.

[85] George Sugihara and Robert M May. Applications of fractals in ecology. *Trends in Ecology & Evolution*, 5(3):79–86, 1990.

[86] Yusak O Susilo and Ryuichi Kitamura. Analysis of day-to-day variability in an individual's action space: exploration of 6-week mobidrive travel diary data. *Transportation Research Record*, 1902(1):124–133, 2005.

[87] Johan AK Suykens and Joos Vandewalle. Least squares support vector machine classifiers. *Neural processing letters*, 9(3):293–300, 1999.

[88] Michael AP Taylor, Jeremy E Woolley, and Rocco Zito. Integration of the global positioning system and geographical information systems for traffic congestion studies. *Transportation Research Part C: Emerging Technologies*, 8(1-6):257–285, 2000.

[89] Arvind Thiagarajan, Lenin Ravindranath, Katrina LaCurts, Samuel Madden, Hari Balakrishnan, Sivan Toledo, and Jakob Eriksson. Vtrack: accurate, energy-aware road traffic delay estimation using mobile phones. In *Proceedings of the 7th ACM Conference on Embedded Networked Sensor Systems*, pages 85–98. ACM, 2009.

[90] Benoit Thierry, Basile Chaix, and Yan Kestens. Detecting activity locations from raw gps data: a novel kernel-based algorithm. *International journal of health geographics*, 12(1):14, 2013.

[91] Caetano Traina Jr, Agma Traina, Leejay Wu, and Christos Faloutsos. Fast feature selection using fractal dimension. *Journal of Information and data Management*, 1(1):3, 2010.

[92] Mariko Utsunomiya, John Attanucci, and Nigel Wilson. Potential uses of transit smart card registration and transaction data to improve transit planning. *Transportation research record*, 1971(1):118–126, 2006.

[93] Rohit Verma, Surjya Ghosh, Niloy Ganguly, Bivas Mitra, and Sandip Chakraborty. Smart-phone based spatio-temporal sensing for annotated transit map generation. In *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, page 16. ACM, 2017.

[94] Yi Wang, Jialiu Lin, Murali Annavaram, Quinn A Jacobson, Jason Hong, Bhaskar Krishnamachari, and Norman Sadeh. A framework of energy efficient mobile sensing for automatic user state recognition. In *Proceedings of the 7th international conference on Mobile systems, applications, and services*, pages 179–192. ACM, 2009.

[95] Jun Yang. Toward physical activity diary: motion recognition using simple acceleration features with mobile phones. In *Proceedings of the 1st international workshop on Interactive multimedia for consumer electronics*, pages 1–10. ACM, 2009.

[96] Robert S Yuill. The standard deviational ellipse; an updated tool for spatial description. *Geografiska Annaler: Series B, Human Geography*, 53(1):28–39, 1971.

[97] Shannon N Zenk, Amy J Schulz, Stephen A Matthews, Angela Odoms-Young, JoEllen Wilbur, Lani Wegrzyn, Kevin Gibbs, Carol Braunschweig, and Carmen Stokes. Activity space environment and dietary and physical activity behaviors: a pilot study. *Health & place*, 17(5):1150–1161, 2011.

[98] Rui Zhang, Kevin G Stanley, Scott Bell, and Daniel Fuller. A feature set for spatial behavior characterization. In *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 512–515. ACM, 2018.

[99] Hongzi Zhu, Yanmin Zhu, Minglu Li, and Lionel M Ni. Hero: online real-time vehicle tracking in shanghai. In *INFOCOM 2008. The 27th Conference on Computer Communications. IEEE*, pages 942–950. IEEE, 2008.