I will follow the following structure to write the paper. And I will fill each part one by one.

Some parts like literature review have been filled, but it is the first version so that still need to be edited.

**Title**

Brief and topic-specific summary of the whole research

**Abstract**

A concise summary of the content and direction of the paper. I will write this part last.

**Introduction**

The introduction will be split to two parts: motivation and background.

1. **Motivation**

This part will mainly introduce why I am undertaking this research. I will explain it from two aspects: 1. The reason why I am interested in this topic (the outbreak of COVID-19 in my country China); 2: The reason why this topic worth investigating (a lot of benefit for finding out the relationship between movement data and the outbreak of infections).

1. **Background**
2. The context of the research (the rationale for the present study)
3. What the research is about (an outline of the research questions and hypotheses)
4. The plan of doing the research
5. A brief introduction of methods used in latter research
6. Key terms and definitions (terminology)

In recent years, due to the population explosion and the convenient of transportation, the incidence of infectious diseases shows a sharp increasing trend. And it becomes a major public health problem globally, killing more than 13 million people each year (Sorell, 1985). Several studies, particularly from rural areas, have shown a link between mobility and HIV infection (Lydié et al., 2004). As the mobility of people who with the infectious virus, infectious diseases can be spread to a wider area. In this case, it is possible that mobility data of population in a certain area contains information about the spatial distribution or outbreak trend of infection. Mobility data describes the mobility behaviour of people in a particular area, such as the journey-to-work data in a country which basically describes the trajectory of people going to work at a country scale. If we can find the relationship between the outbreak trend of infectious diseases and mobility data through the establishment of models, then we can take certain measures to prevent the outbreak or spread up to a point. Moreover, Modeling is the process of simulating the causal relationship between two things. In order to reveal the relationship between mobility data and the distribution of infectious diseases, the study is divided into two stages: mining mobility data; modeling and analyzing.

In the past year, with the unstoppable global outbreak of COVID-19, many countries have adopted travel bans to prevent the further spread of infectious diseases. Undoubtedly, the movement of people will accelerate the transmission of infectious diseases. However, mobility data describes the mobility behaviour of people in a particular area, such as the journey-to-work data in a country. Otherwise, not only efffectively reflects human behavior and activities, movement data also contains a wealth of information on user behavior patterns and important location attributes. Due to the development of mobile application technology and spatial data acquisition technology, it is not difficult to obtain user’s trajectory data. Zheng (2015) proposed the entire process of trajectory data mining as well as techniques applied in each step [2]. In addition, social interactions are another concerned factor that promotes the spread of diseases. Past research has investigated the attributes of social encounters by conducting a telephone survey [3].

Moreover, great understanding of the relationship between infectious disease transmission and the trajectory of human movement is critical for predicting and preventing future outbreaks of infectious diseases. And, effective public interventions can be designed in a timely manner to control the further spread of infectious disease in this case [1]. Much research has been done to explore the association between the transmission of infectious diseases and geographical or demographic indicators (such as per-head income) at different scale (city, urban, country etc.) [1,4]. Furthermore, the research incorporating trajectory data with social encounters have been conducted to reveal the spatial distribution of infection depending on specific locations [6].

Here, we will combine the trajectory data with social encounters to explore the spatial distribution of infectious disease. In this process, cluster algorithms, which are a group of efficient algorithms to do unsupervised machine learning, will be applied to find general patterns in movement dataset. Then, the social interaction of participants and basic information (such as age etc.) will be combined with movement patterns to reveal the law of disease distribution.

Mobility data describes the mobility behaviour of people in a particular area, such as the journey-to-work data in a country. If we can accurately find the relationship between the outbreak trend of infectious diseases and mobility data through the establishment of models, then we can take certain measures to prevent the outbreak. Much Research has been done to identify the model of the spread of infections based on mobility data at either national and international or individual buildings such as school or hospitals.

Moss et al focus on the infection at metropolitan scale. However, how to gain appropriate mobility data increasingly becomes a difficult and important topic. Many technologies have been used to mining this kind of data. Zheng et al. developed a complete process to mining and process trajectory data of human or animals.

But there are still two kinds of challenges: 1. Collecting data; 2. Analyzing data. About collecting data, some important data is hard to collect, such as those characterise private features. When it comes to analyze the data, basicly, even just applying corresponding method to a particular data set is a big challenge, and combining mobility data and social encounters of respondents is necessary, because some behaviors may have effect on infection. Moreover, Roll et al. analyze social encounter of some respondents, and give results based on age, gender and so on.

This research will focus on using existing methodologies and algorithms to mine new trajectory dataset at urban scale that consider the factors that may affect the chance of infecting, such as the length for a person staying at one place and so on. Also, based on what Roll et al have discovered, this research will identify the technologies that can combine social encounter data and mobility data, and use it to reveal about the pattern or predict the trend of infection.

1. Chowell G, Bettencourt LMA, Johnson N, Alonso WJ, Viboud C. The 1918–1919 influenza pandemic in England and Wales: spatial patterns in transmissibility and mortality impact. Philos Trans R Soc B Biol Sci. 2008;275(1634):501–9. <https://doi.org/10.1098/rspb.2007.1477>.
2. Zheng, Y. (2015). Trajectory data mining: an overview. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *6*(3), 1-41.
3. Rolls, D. A., Geard, N. L., Warr, D. J., Nathan, P. M., Robins, G. L., Pattison, P. E., ... & McVernon, J. (2015). Social encounter profiles of greater Melbourne residents, by location–a telephone survey. *BMC infectious diseases*, *15*(1), 1-11.
4. Murray, C. J., Lopez, A. D., Chin, B., Feehan, D., & Hill, K. H. (2006). Estimation of potential global pandemic influenza mortality on the basis of vital registry data from the 1918–20 pandemic: a quantitative analysis. *The Lancet*, *368*(9554), 2211-2218.
5. Viboud, C., Bjørnstad, O. N., Smith, D. L., Simonsen, L., Miller, M. A., & Grenfell, B. T. (2006). Synchrony, waves, and spatial hierarchies in the spread of influenza. *science*, *312*(5772), 447-451.
6. Moss, R., Naghizade, E., Tomko, M., & Geard, N. (2019). What can urban mobility data reveal about the spatial distribution of infection in a single city?. *BMC public health*, *19*(1), 1-16.

**Literature review**

The review will be a selection of carefully organized, focused and relevant literature that develops a narrative ‘story’ about the topic. The review will answer questions about the literature:

1. What is the current state of knowledge on the topic?
2. What differences in approaches / methodologies are there?
3. Where are the strengths and weaknesses of the research?
4. What further research is needed?
5. Identify a gap in the literature which provides a rationale for my study and supports the research questions and methodology.

1. mobility data mining

With the maturity of location acquisition technology, the popularization of wireless network and the convenience of transportation, a large number of trajectory data created by human, animal, transportation tools and even climate change have been excavated for making applications or doing research (Zheng, 2015). In “Trajectory Data Mining”, Zheng proposes the whole process of mining and processing trajectory data created by moving objects (Colyer, 2019), which is mainly composed of the derivation of trajectory data, trajectory data preprocessing, trajectory data management, and various data mining tasks. Zheng also introduces the similarities and differences between the existing technologies of mining data. In addition, this paper also identifies methods of transforming trajectory data into other data structures, such as graphics, matrices and tensors. Data in these formats can be applied to more apps and researches, which facilitates this research topic. At the end of the article, the author presents some trajectory data sets for use by researchers or developers. Also, it points out the future direction of research based on this data set (Zheng, 2015). The trajectory data mining technology proposed by Zheng encourages researchers and developers to research or develop applications based on the analysis of trajectory data. In addition, the data set provided in this paper facilitates the subsequent study of the spatial distribution of infectious diseases.

However, it involves many problems and challenges that must be met during the data mining process. Firstly, in many countries and regions, obtaining location and mobility information on mobile phones remains difficult and even restricted by legislation (Wesolowski, Buckee, Engø-Monsen & Metcalf, 2016). Moreover, despite the high prevalence of mobile phones, the data can be biased. For instance, without possession of mobile electronic devices, the trajectory data of children generally is not recorded. Therefore, the main source of data is adults. But for infectious diseases, children are also a major source of infection. Another point is the collection of data can be affected by making or receiving calls. Although the impact is small, it involves the privacy concerns of device owners. Furthermore, it is difficult to detect position of device owners due to the weakness of signal in some particular area. In this case, the trajectory is not continuous which lead to the inaccuracy of the analysis of spatial distribution of infections.

According to Hung, Peng and Lee (2011), there are specific solutions to solve the last challenge. Generally, there are many trajectories in a particular area created by different people,

Generally, some clues can be found from these trajectories. So, from we can cluster the trajectories based on these clues. Hung et al. call this approach the clue-aware clustering algorithm (CACT). In a hot spot, there must be some trajectories share the same clues. For example, following a prescribed touring route in a scenic spot, all the visitors create the same trajectory. In this case, this scenic spot is a clue that can help cluster trajectories. Based on this trajectory group and CACT, we can find the common track of people in a hot spot even if the mobile device only captures some motion fragments (Hung, Peng & Lee, 2011). CACT solves the potential problems caused by incomplete trajectories mentioned in ‘Trajectory Data Mining’ and further improve the process of data mining. However, a lot of information other than track is needed, and simple stay point detection is not sufficient for clustering trajectories.

2. modeling and analyzing

Much Research has been done to identify the model of the spread of infections based on mobility data at either national and international or individual buildings such as school or hospitals. However, in fact, the mobility pattern at city scale is quite different from country scale (Moss, Naghizade, Tomko & Geard, 2019). According to Moss, Naghizade, Tomko and Geard (2019), the most appropriate size of mobility data is that collected at metropolitan scale. In this article, the author conducts the research form determining the data set of the study (journey-to-work and GPS data set) based on the city size, to analyze the influence of hub and spoke commuting patterns, to apply mathematical method to represent the mixing pattern, to establish model, finally analyze the relationship between the spatial distribution of infectious diseases and the mobility data. Also, the article identifies some problems to be solved. Privacy issues are inevitably when doing such research, which makes it difficult to obtain report of infectious cases. Secondly, it is also a major challenge to combine case information with mobility information. In all, Moss et al. innovatively propose to explore the spatial distribution of infectious diseases from the perspective of cities. Although the mobility and contact of people in urban areas are more complex and difficult to identify, it provides some ideas and methods for studying the distribution of infectious diseases in urban areas. For example, consider the impact of the characteristics of highly connected areas, or focus on the information contained at transportation hubs.

Due to the lack of consideration of contact information between people and the focusing on the mobility information, there may be some bias for the conclusion. In other words, some uncertainties may affect the results of the analysis, such as the duration that an object spends at a certain location, or the influence of the heterogeneity of people's social encounters. Rolls et al. (2015) conduct a telephone survey on the social encounter attributes of people in two areas of Melbourne. It turns out that regardless of gender, age, location and area of residence, people's contact behavior patterns are completely different. For example, the frequency and length of contact between adult women and young children is higher than that of adult men of the same age, which may explain the fact that young children are more likely to be cared for by their mothers from a sociological perspective. In this case, if one of the members in this family carry infectious virus, other members in the family can be infected. In addition, this paper identifies several kinds of human-to-human interactions that may affect the distribution of infectious diseases. Including duration of social encounters, encounter with known or unknown individuals, the impact of local government area of residence. And it also gives conclusions like “The highest reported number of median contacts was among individuals aged between 30 and 49 years”. Assumptions about the number, duration, and aggregation of contacts in the model framework all have important implications for the simulated transmission of infection, without building model based on this data set and analyzing the concrete relationship between social encounters and the reported disease cases, this article cannot adequately explain the point that social encounter can affect the distribution of infection.

Other technologies have been proposed to extract information from dataset, by combining trajectory sample points with geographic data, behavioral knowledge which is more useful and meaningful to application users can be extracted from semantic trajectory (Chakri, Raghay & El Hadaj, 2017). And Mossong et al conduct a survey to record and analyze different contact patterns of different individuals during one day. And give conclusion that 5- to 19-year-olds are expected to suffer the highest incidence during the initial epidemic phase of an emerging infection transmitted through social contacts measured here when the population is completely susceptible (Mossong et al., 2008). It is the first paper that provides quantitative approaches to study the contact patterns of infectious disease.

3. Challenge

Although much research has been done for revealing the laws of the distribution of infection, there are still some challenges. From collecting data to processing data to analyzing data, there are unsolved problems.

Firstly, for collecting data, new approaches in recent years have given us the opportunity to collect data on individuals and population sizes to empirically describe patterns of exposure within host populations (Eames, Bansal, Frost & Riley, 2015). But there are still some challenges. It is divided into two parts: mobility data collecting, social encounter profile identifying. In terms of mobility data collecting, it is hard to decide which scope is most appropriate. Sometimes, it also involves privacy problems. When it comes to social encounter data collecting, six challenges presented by Eames, Bansal, Frost & Riley, (2015). It involves the difficulty of defining a contact and avoiding privacy issues such as personal information and ethical concerns (Eames, Bansal, Frost & Riley, 2015). In particular, some data collection processes may involve illegal content even though it is for research purpose.

The next part is about processing data. Quantifying data can be a big challenge. As the data obtained from such surveys are generally descriptive, it is difficult to use some numerical methods to represent it. Since no matter how you represent this kind of data, not all the information is going to be contained. Other challenges of processing data are proposed by Eames et al., including bounding networks in space, time, and scope, dealing with missing data, measuring weighted and dynamic networks, exploiting in direct information about networks (Eames, Bansal, Frost & Riley, 2015).

Moreover, in order to comprehensively analyze the distribution of infectious diseases and predict the incidence trend of infectious diseases, many factors need to be considered. Firstly, incorporating geographical space into representations of social networks in the field of infectious disease modelling is one of the biggest challenges (Rolls et al., 2015). Moreover, infections are always seasonal, so time is also an important factor that can affect the distribution of infections, similar factors are climate, temperature and so on. In addition to these environmental factors, proper techniques for extracting information from data sets should be figured out. What's more complicated is that conclusions based on a particular region are not universal. In other words, every country, region or city is of different situation. The law cannot be applied to predict infection trend of other cities or regions.

After obtaining proper dataset, some technologies should be applied to build model. Traditionally, the distribution of infectious diseases in human populations has been modelled with static parameters (Funk et al., 2015). But with individuals changing their behaviour, these parameters can change. In other word, how to incorporate these behavioural changes in models of infectious disease is the challenge.

Summary of literature

1. It explores the association between influenza death rates, transmissibility and several geographical and demographic indicators for the autumn and winter waves of the 1918–1919 pandemic in cities, towns and rural areas of England and Wales. It finds that death rates varied markedly with urbanization, and there is no association between transmissibility, death rates and indicators of population density and residential crowding [1].
2. In this paper, the association between per-head income and mortality rate during influenza pandemic are examined. It shows high mortality rate in poor country [2].
3. In this paper, the spread of infectious disease between state is evaluated. It shows that the regional spread of infection correlates more closely with rates of movement of people to and from their workplaces (workflows) than with geographical distance.
4. It gives the definition of trajectory and supplies several improved cluster algorithms for clustering trajectories. Due to the property of varying lengths of trajectory data, some cluster algorithms are improved in order to apply on trajectory data. Also, Some metrics are provided to evaluate the algorithms.
5. This paper focuses on such a nature of human movements as a trajectory in

two or three dimensional spaces and proposes a method for grouping trajectories

as two-dimensional time-series data, consisting of the following two steps. Firstly,

it compared two trajectories based on their structural similarity, determines the best

correspondence of partial trajectories and calculates the dissimilarity between the

sequences. Then clustering method are applied by using the dissimilarity matrix. The method illustrated in this paper is not applicable of dataset in this research. Because it emphasizes the spatial information contained in trajectory data. But it offers a way to do clustering. And the concept of dissimilarity matrix is valuable to be trying.

1. This paper illustrates the importance of temporal data mining. In terms of clustering, it provides two classes of commonly used methods: model-based and alignment-based methods. In model-based method group, HMM is good at dealing with sequence input. Alignment-based methods are cluster trajectories based on the distance between trajectories.
2. It is not always helpful to explore patterns in the entire sequence. Some important information often contained in segment of trajectory sequence. This paper proposed Partition and group based clustering method to find patterns by segmenting the whole trajectory.

[1] The 1918–1919 influenza pandemic in England and Wales: spatial patterns in transmissibility and mortality impact.

[2] Estimation of potential global pandemic influenza mortality on the basis of vital registry data from the 1918–20 pandemic: a quantitative analysis

[3] Synchrony, Waves, and Spatial Hierarchies in the Spread of Influenza

Reference

[4] Bian, J., Tian, D., Tang, Y., & Tao, D. (2018). A survey on trajectory clustering analysis. *arXiv preprint arXiv:1802.06971*.

[5] Tsumoto, S., & Hirano, S. (2009). Behavior grouping based on trajectory mining. In *Social Computing and Behavioral Modeling* (pp. 1-8). Springer, Boston, MA.

[6] Laxman, S., & Sastry, P. S. (2006). A survey of temporal data mining. *Sadhana*, *31*(2), 173-198.

[7] Lee, J. G., Han, J., & Whang, K. Y. (2007, June). Trajectory clustering: a partition-and-group framework. In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data* (pp. 593-604).

Sorell, T. (1985). Changing Patterns of Infectious Disease. Pathology, 17(4), 681. doi: 10.1016/s0031-3025(16)36908-2

Lydi??, N., Robinson, N., Ferry, B., Akam, E., De Loenzien, M., & Abega, S. (2004). Mobility, Sexual Behavior, and HIV Infection in an Urban Population in Cameroon. JAIDS Journal Of Acquired Immune Deficiency Syndromes, 35(1), 67-74. doi: 10.1097/00126334-200401010-00010

Zheng, Y. (2015). Trajectory Data Mining. ACM Transactions On Intelligent Systems And Technology, 6(3), 1-41. doi: 10.1145/2743025

Colyer, A. (2019). Trajectory Data Mining: An Overview. Retrieved 3 September 2019, from https://blog.acolyer.org/2016/03/07/trajectory-data-mining-an-overview/

Chakri, S., Raghay, S., & El Hadaj, S. (2017). Semantic Trajectory Knowledge Discovery: A Promising Way to Extract Meaningful Patterns from Spatiotemporal Data. International Journal Of Software Engineering And Knowledge Engineering, 27(03), 399-421. doi: 10.1142/s0218194017500140

Mossong, J., Hens, N., Jit, M., Beutels, P., Auranen, K., & Mikolajczyk, R. et al. (2008). Social Contacts and Mixing Patterns Relevant to the Spread of Infectious Diseases. Plos Medicine, 5(3), e74. doi: 10.1371/journal.pmed.0050074

Amy Wesolowski, Caroline O. Buckee, Kenth Engø-Monsen, C. J. E. Metcalf, Connecting Mobility to Infectious Diseases: The Promise and Limits of Mobile Phone Data, The Journal of Infectious Diseases, Volume 214, Issue suppl\_4, December 2016, Pages S414–S420, https://doi.org/10.1093/infdis/jiw273

Hung, C., Peng, W., & Lee, W. (2011). Clustering and aggregating clues of trajectories for mining trajectory patterns and routes. The VLDB Journal, 24(2), 169-192. doi: 10.1007/s00778-011-0262-6

Moss, R., Naghizade, E., Tomko, M., & Geard, N. (2019). What can urban mobility data reveal about the spatial distribution of infection in a single city?. BMC Public Health, 19(1). doi: 10.1186/s12889-019-6968-x

Rolls, D., Geard, N., Warr, D., Nathan, P., Robins, G., & Pattison, P. et al. (2015). Social encounter profiles of greater Melbourne residents, by location – a telephone survey. BMC Infectious Diseases, 15(1). doi: 10.1186/s12879-015-1237-9

Eames, K., Bansal, S., Frost, S., & Riley, S. (2015). Six challenges in measuring contact networks for use in modelling. Epidemics, 10, 72-77. doi: 10.1016/j.epidem.2014.08.006

Funk, S., Bansal, S., Bauch, C., Eames, K., Edmunds, W., Galvani, A., & Klepac, P. (2015). Nine challenges in incorporating the dynamics of behaviour in infectious diseases models. Epidemics, 10, 21-25. doi: 10.1016/j.epidem.2014.09.005

**Data set**

We explores movement patterns based on the data set “day\_in\_the\_life”. This data set is from a telephone survey conducted by Rolls et al. (2015). R is applied to make the calculation and virtualization of the data set. It is a language and environment to make statistical analysis. Graphs are given to enhance the analysis. Box plots are used to analyze the distribution of features in the data set, such as age etc. It can clearly reflect the scope of the concentrated distribution of features as well as outliers. Proportion bar charts will also be widely used in this paper. In this type of graph, based on one feature staying unchanged, the proportion of different values of another feature will be reflected. This is conducive to horizontal and vertical comparison and analysis.

The original data set contains information such as the age and gender of the respondent, the specific location that the respondents have visited during the survey time period, the type of the specific location (home, work, etc. we call it location type in later parts), arrival and departure time from a certain location, etc. Amount 1307 respondents in the dataset, there are 514 males and 793 females. In addition, with age of -2, error exists in 33 pieces of data. so they are excluded when analyzing the data set.

In terms of age, Figure 1 shows the distribution. Obviously, the age of the respondent is concentrated between 45-70. It means the result is more representative amount people at age of 45-70 rather than every age group. There is no outlier in the figure, . The value of the outlier is -2. Obviously, it is not reasonable to be at age of -2. It should be deleted when it involves in analyzing the association between age and other featuresA screenshot of a social media post

Description automatically generated

A screenshot of a cell phone

Description automatically generated

Figure 1: The distribution of age

Although age is a discrete variable, it is too trivial to do the analysis based on each specific age. Rolls et al. (2015) categories the participants into five groups based on age. Regardless of -2 and “refused” data, they are categorized as 18-29, 30-49, 50-59, 60-69, 70+ years, respectively. Some combination of age group are made by weighting to address the problem of age related bias. The proportion of participants in each group are given in Table 1:

|  |  |  |
| --- | --- | --- |
| Initial categories | 5 category for weighting | Proportion |
| 18-19 | 18-29 | 0.098 |
| 20-29 |
| 30-39 | 30-49 | 0.217 |
| 40-49 |
| 50-59 | 50-59 | 0.221 |
| 60-69 | 60-69 | 0.223 |
| 70 or more | 70 + years | 0.240 |

Then, the gender is considered. The number of female participants is a little bit more than male participants. But as you can see in Figure 2, the distribution of participants with different age in two gender groups are almost the same.

A screenshot of a cell phone

Description automatically generated

Figure 2: Age distribution based on gender. 1 represents male and 2 represents female.

In this paper, we care much about where people have been in a time period. We start from how many locations respondents have been. Then the number of unique locations are considered. Longitude and latitude are compared to determine whether it is the same location. In other words, if two places have the same latitude and longitude pairs, they are the same place, and count as one, vice versa. As shown in Figure 3, in general, people visit 4-10 locations in a day but 40 places at most. In this perspective, it is too trivial to make the analysis on the basis of the number of locations.

A screenshot of a cell phone

Description automatically generated

Figure 3: The distribution of number of locations

Therefore, location types are introduced to simplify and structure the problem. Figure 4 shows the distribution of the number of unique location types visited. There are totally 13 location types. Classifying different places as the same location type, it is much less trivial than considering unique places as well as helpful for exploring movement patterns. Most participants visit places of only 2 to 4 types. This can be interpreted from two perspectives. Roughly group all of the participants as 18-59 and 60 + years according to whether they have retired or not. For group 18-59, the trajectory pattern in the working day is almost fixed, or similar, which is between home and workplace. As for group with age more than 60, they prefer staying at a place for long time rather than moving frequently.

A close up of a logo

Description automatically generated

Figure 4: Distribution of the number of unique location types

Then, the distribution of locations or location types will be combined with the basic information (age, gender) of participants to discover basic rules hidden in the dataset which can help analyze the result in the latter step.

Firstly, we try to find the relationship between the number of locations and the age of the respondents. When referring to the number of locations, it means the number of locations visited by the respondent within the time period of the survey. When the respondents repeatedly arrive at a place, it is counted twice. As shown in c in Figure 5, without considering no-age group, as age increases, the proportion of high-frequency movement gradually decreases. However, the age group of 18-29 is an outlier of this trend. Moreover, in every age group, 5 places usually accounted for the largest proportion while two locations hardly appear in any age group. Imagine a participant going out home for something, he must return home at the end of the day and there must be a third place between two homes. Thus, it should be either 1 or ≥3. Sometimes, a place (such as home) can be recorded twice even if the participant stays there all the time but the longitude and latitude changed .

A screenshot of a cell phone

Description automatically generatedA picture containing implement, stationary, pencil, colorful

Description automatically generated

1. b.

A picture containing drawing

Description automatically generatedA close up of a logo

Description automatically generated

c. d.

Figure 5: age range and the number of places

Next, the following figures shows the relationship between the age of participants and the number of unique locations the participants have been to during 11 hours. Different from the above statistics, the repeat place will be counted once. In this dataset, longitudes and latitudes of specific places are provided except for movement type. Having excluded all the movement locations, such as Private Transport, all the repeat places can be eliminated by comparing the longitude and latitude between different locations. The two figures in first column reveal the distribution of the number of unique locations on the basis of age group. Here, the meaning of the number one needs to be clarified. Generally speaking, without considering the error caused by the survey method and movement locations, participants will go to at least two different types of places, which are home and one other places (Work, Study or something like this). Therefore, one can only be interpreted as that the participant does not go out during the survey period. In other words, the participant has been in the same location without moving. And this type of location is most likely home. This also proves that in Figure 1, participants aged 70+ have the largest proportion of participants who have visited one place. Due to retirement or healthy problems, they stay at home all day. Thus, it is only one place visited. In every age group, 3 accounts for the largest proportion.

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generatedA picture containing implement, stationary, pencil

Description automatically generated

A picture containing drawing

Description automatically generatedA close up of a logo

Description automatically generated

Relationship between age group and location types

Generally speaking, it is too trivial to take specific location into consideration. Instead, location type is what worthwhile to focus on. There are totally 12 location types. According to the results generated by R, the maximum number of location types visited is 9. For participants who have visited many types of places, instead of focusing on specific frequency, we care more about the proportion. Frequencies above six are referred to high frequencies, so they are all represented by “6+” in plot. As shown in the left two plots in Figure below, with the highest proportion amount each age group, 3 is interpretable when the transportation between locations is considered. For instance, three locations types can be home, a kind of transportation (private transport, etc.), work (or study, etc.). Moreover, same as the trend in locations and age group, age is inversely proportional to the number of location types except for 18-29 group.

A close up of a piece of paper

Description automatically generatedA picture containing implement, stationary, pencil

Description automatically generated

A picture containing drawing

Description automatically generatedA picture containing drawing

Description automatically generated

Relationship between gender and number of locations

In terms of gender, the distribution of locations is almost the same.

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated A picture containing drawing

Description automatically generated

A picture containing large

Description automatically generated A screenshot of a cell phone

Description automatically generated A picture containing drawing

Description automatically generated

Duration refers to the total length of time the participant stayed in a certain type of place. Through the entire activity track, participants will stay in the same type of place for discrete periods of time. In this way, the sum of all the inconsecutive interval represents the duration the respondent stayed in the corresponding type of place. As shown in the figure, the medians of “Home” and “Work” are higher than other types of locations. Since they all represent the mode of transportation, the distributions of private transport, public transport and car journey are basically the same. Far fewer time spent in transportation, they are all recategorized as “travel”. Likely, “AC”, “PS”, “SR” are set as “entertainment”. The distribution after recategorizing are shown in the second figure.

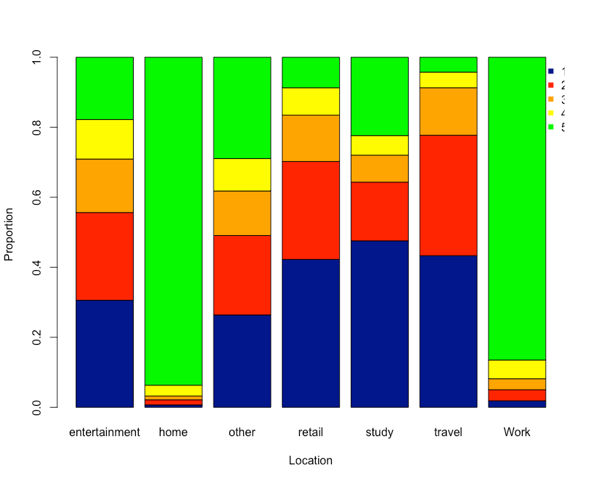
A close up of a map

Description automatically generatedA screenshot of a cell phone

Description automatically generated

To simplify the process, we divide the duration into 5 groups. Grouped into a group every hour. The numbers 1, 2, 3, and 4 do not represent the exact number of hours, but rather the duration range. For example, 1 represents any time interval between 0 minute and 60 minutes, which can be 30 minutes or 45 minutes, etc. For each duration group, we counted the proportion of people staying in different types of places. With the increasing of the length of time, the proportion of participants who stay at home increases as well. Lower proportion than “home” in “5+” group, the proportion of work gradually increases with the length of time becoming longer. This shows that people prefer to spend a large amount of time at home and a small amount of time at entertainment, retail etc. In addition, the elderly respondents accounted for the majority of the survey population, so larger amount of time was spent at home rather than at work. Next, comparing the duration distribution of participants of different ages and genders is helpful for analyzing the result of clustering in later steps.

A close up of a logo

Description automatically generated

For different genders, the distribution of time spent in different types of places is roughly the same. As discussed above, there is almost no difference between the age distribution amount female and male. However, the main factor leading the difference in duration distribution is age. In other words, each age group has its specific trajectory pattern and duration distribution, thus if the age distribution of participants of different genders is roughly the same, then it is reasonable that the distribution of duration between genders is the same. This can be proved in the following comparison of the duration distribution between different ages.

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

Still follows the previous age groups, for some type of locations, the distribution in each age group is quite different from each other. Take home as an example, except for the -2 age group, the length of time the participants stay at home gradually increases as the age increases. In addition, the distribution of study time varies for each age group. Same as “home”, with the increase of age, the time of study is on the rise. But there is an abnormality. That is, the study time of participants aged 18-29 is higher than that of other age groups. Generally, participants aged 18-29 should be in school, and they will spend more time studying. So it is reasonable.

A screenshot of a social media post

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A picture containing screenshot

Description automatically generatedA picture containing screenshot

Description automatically generatedA screenshot of a cell phone

Description automatically generated

**Methodology**

The purpose of this section is to detail how I conducted my research so that others can understand and replicate my approach. I will briefly describe the equipment or materials used and the approach taken. There will be two parts in this section.

1. Methods(techniques) used in Trajectory patterns exploration

Need more research

Mobility data describes the mobility behaviour of objects in a particular area, such as the journey-to-work data in a country. Movement patterns can be explored from mobility data. Having known patterns of movement, we can discuss the relationship between the outbreak of specific infectious disease and movement patterns. Then, prevent the outbreak by predicting the trend of outbreak.

Much Research has been done to identify the model of the spread of infections based on mobility data at either national and international or individual buildings such as school or hospitals. Moss et al. focus on the infection at metropolitan scale. However, how to gain appropriate mobility data increasingly becomes a difficult and important topic. Many technologies have been used to mining this kind of data. Zheng et al. developed a complete process to mining and process trajectory data of human or animals.

To apply the techniques on the dataset, the representation of dataset needs to be figured out. We focus on location type that one has been rather than specific location. There are totally 12 location types: Home, Public transport, private transport, Retail and hospitality, Car journey, Sport and recreation, Public spaces, Work, Study, Arts and culture, Other, Refused. Refused means respondents does not tell researcher about their movements that day, and it is eliminated from the sequence. Therefore, the length of full movement trajectory will be 11. The followings are some representations used in this paper.

1. Either a participant has been to a location or not. 0 represents a participant has not been to the location type while 1 suggests the participant has been to the corresponding location type. A piece of data is represented as a vector of zero and one with length of twelve. The full sequence will be a vector of twelve 1s. In this way, a predefined sequence of location type is required, which is ['Arts and culture', 'Car journey', 'Home', 'Other', 'Private Transport', 'Public Transport', 'Public spaces', 'Refused', 'Retail and hospitality', 'Sport and recreation', 'Study', 'Work']. To interpret an vector, the result of performing ‘AND(∧)’ logic operation are the location types the participant has been in a period. In other words, we do not care about the order information, but it significant to provide a predefined order to unify the description of every piece of data.
2. How many times a participant has been to a location type. Same as the first type of representation, each piece of data in this representation is also a vector of length twelve. However, instead of 0 or 1, each element of this vector is an arbitrary number which indicates the number of times the participant has visited the corresponding location type during the survey time.
3. How long a participant has been at a location type. In this dataset, the exact time of arriving at a location and the departure time are given. By making calculation from the two time, we can get the duration a participant has stayed in a place. There are two variations: 1. As the last representation, it is also a vector with length of eleven, but the element will be duration. If one has been to a location type twice in a day, just add the duration up; 2. Add order information in representation. We do not add duration up. Such as the table below, it can be represented as (1,8,10).

|  |  |  |  |
| --- | --- | --- | --- |
| Location type | Home | Work | Home |
| Duration (hours) | 1 | 8 | 10 |

1. Instead of representing the trajectory in numeric format, text format loses less information. Thus, a trajectory is represented as length-variation vector of text. There are two variations of this representation: 1. The order of locations in which the respondents been is considered here. And the repetition of location types is all kept. For example, (Home, Work, Private transport, Home) is different from (Home, Private transport, work, Home), and they are all valid; 2. Use different numbers to represent different types of places. In this way, the order information can be maintained as well as it can be in numeric format.
2. Most algorithms only accept input with the same length. To ensure the same length of the input and keep the order information of the trajectory, it is necessary to fill some other values in the input vector. Provided location type that the participant has been from 7:00am to 23:00pm, we firstly set the length of the input vector to 16. Then each element in the vector represents a time from 7 o'clock to 12 o'clock. The value of each element is the location type that the participant stays at the corresponding time. To be simplified, we use number from 1 to 12 or initial letter to represent each location type. However, the drawbacks are obviously. Firstly, instead of staying at one location, people always travel to many places in an hour. It is hard to decide which value to choose. For instance, given the information below, the third value in the vector, which suggests the location type one has been to from 9:00am to 10am, are car journey and sport and recreation. Due to fix-length input is needed, it should keep either car journey or sport and recreation. Compared the durations of two location types, it seems sport and recreation should be the value while car journey needs to be discarded. However, both of them are significant to explore movement patterns.

|  |  |  |
| --- | --- | --- |
| Location type | Time of Arrival | Time of Departure |
| Home | 2013/3/9 7:00 | 2013/3/9 9:00 |
| Car journey (respondent alone in the car) | 2013/3/9 9:00 | 2013/3/9 9:18 |
| Sport and recreation | 2013/3/9 9:18 | 2013/3/9 10:30 |
| Retail and hospitality (bars, cafes, shops, hair dressing, etc.) | 2013/3/9 10:30 | 2013/3/9 11:30 |
| Car journey (respondent alone in the car) | 2013/3/9 11:30 | 2013/3/9 11:48 |
| Home | 2013/3/9 11:48 | 2013/3/9 23:00 |

Machine learning techniques

1. Naïve bayes.

This method is applied to explore the association between trajectory and gender or age. This method will be applied in three ways: 1. The first two kinds of representation can be treated as feature vector while gender or age is label. When applying this method, the whole dataset will be split to two parts. One is for training the model, another is for evaluating the model; 2. The age range and gender are features while location types are labels. After obtaining the model, the trajectory patterns in terms of age and gender will come out, and this will help analyze the dataset; 3. Age, gender and part of the trajectory are the features, the next location type it will be is treated as label.

1. Global edit distance

This method calculates the distance between trajectories. A threshold of distance needs to be predefined. If the distance between two trajectories exceed the threshold, they belong to two groups. Otherwise, they can be classified in the same group. For example, the distance between (Home, Private transport, Work, Retail and hospital, Private transport, Home) and (Home, Public transport, Work, Public transport, Home) is three. But the drawback is obvious. Take the same example, if the threshold is 2, the two trajectories are different. However, they have the same pattern.

1. Cluster

Cluster is an unsupervised machine learning technique. It can classify trajectories in different groups. Then, the centroid of each group are the patterns we suppose to find. The drawback is the number of centroids is required to be predefined. If the number is not appropriate, the pattern is not representative enough. Some commonly used algorithms of cluster are K-means, Mean-shift, Spectral Clustering and Hierarchical clustering. They can be applied on all of the representations of dataset.

Centroids need to be data points in original dataset; drawbacks of treating centroids as trajectory patterns: cannot make summarization; dataset analysis; HMM

**Results**

In this section, I will show patterns found by each technique in methodology part. It will be the tabular or graphic summary of the findings.

No analysis of these results will be given in this part

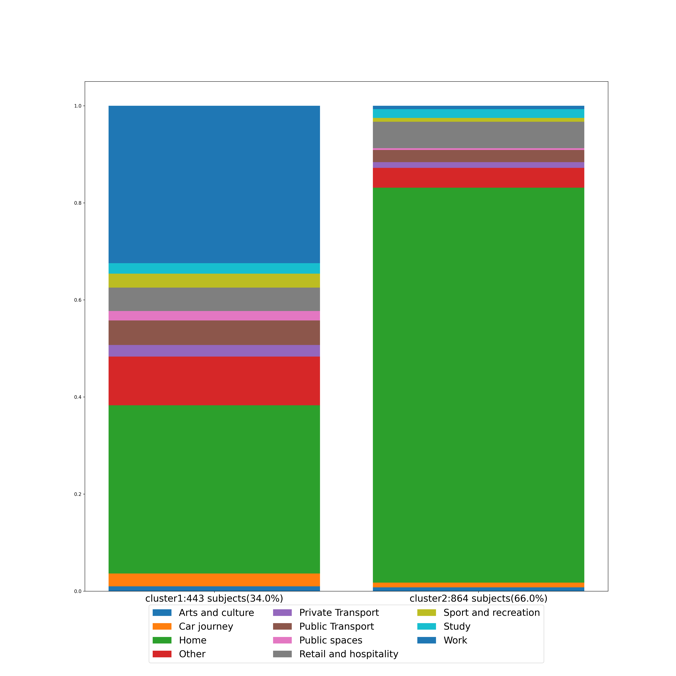
The performance of different cluster methods on different representations of the data set is given in the following three tables. Since no correct label for each piece of data, this is a completely unsupervised learning cluster. Therefore, we use SC, CHI, and DBI to measure the performance of each combination of cluster algorithm and the vectorized representation of different trajectories.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Duration | | | (01) | | | Frequency | | |
|  | SC | CHI | DBI | SC | CHI | DBI | SC | CHI | DBI |
| Kmeans(n=4) | 0.431 | 871.881 | 1.267 | 0.229 | 214.259 | 1.815 | 0.257 | 333.754 | 1.416 |
| Mean-shift | 0.418 | 183.441 | 1.034 | 0.164 | 25.872 | 1.998 | 0.238 | 36.395 | 1.133 |
| Spectral Clustering (n=4) | - | - | - | 0.179 | 179.635 | 1.904 | 0.499 | 81.373 | 0.895 |
| Hierarchical clustering | 0.470 | 813.645 | 0.778 | 0.121 | 142.416 | 2.596 | 0.398 | 309.942 | 1.249 |
| DBSCAN | -0.301 | 8.881 | 1.030 | 0.253 | 2.568 | 1.790 | -0.244 | 8.629 | 1.490 |
| OPTICS | 0.142 | 7.482 | 1.215 | 0.866 | 58.826 | 1.057 | 0.326 | 6.868 | 1.334 |

Duration

|  |  |  |  |
| --- | --- | --- | --- |
|  | SC | CHI | DBI |
| Kmeans(n=2) | 0.474 | 987.210 | 1.043 |
| Kmeans(n=3) | 0.475 | 1022.566 | 1.064 |
| Kmeans(n=4) | 0.431 | 871.881 | 1.267 |
| Kmeans(n=5) | 0.398 | 779.455 | 1.381 |
|  |  |  |  |

Kmeans with different number of clusters (duration)

A picture containing drawing, room

Description automatically generated

A picture containing drawing

Description automatically generatedA picture containing drawing

Description automatically generated

Meanshift after eliminating clusters with subjects <= 50

A picture containing drawing

Description automatically generated

A picture containing food

Description automatically generated

Figure: meanshift average

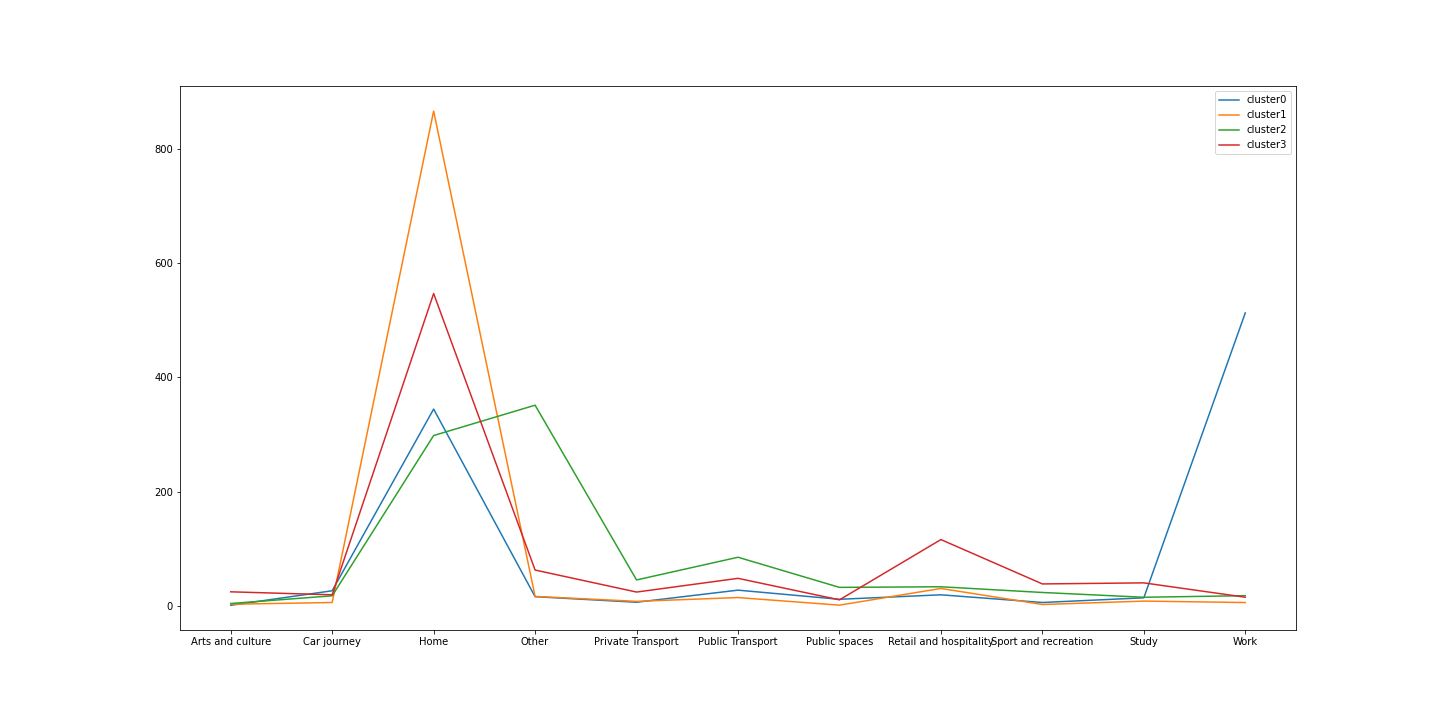


Figure: duration and location type

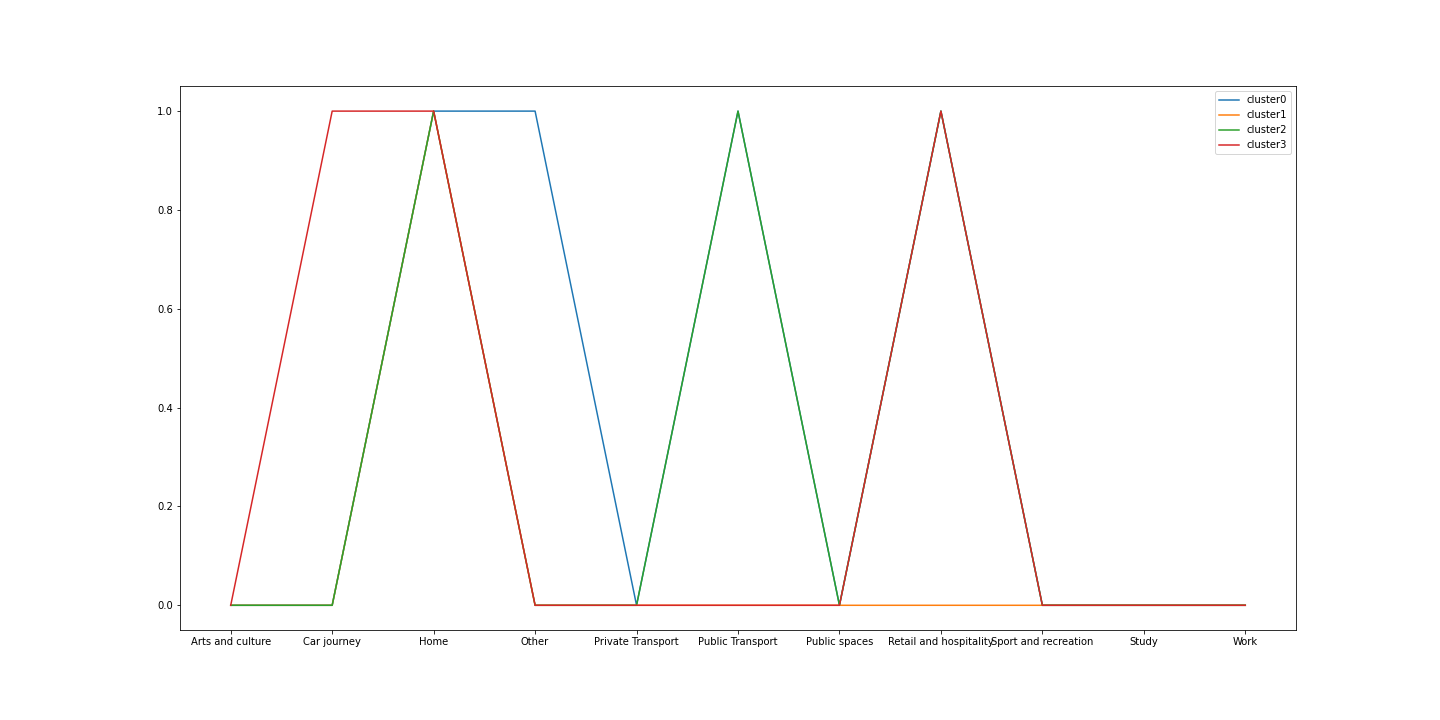
A picture containing sitting

Description automatically generated

A picture containing sitting

Description automatically generated

Figure: frequency



**Discussion**

Combined with three representations of the dataset, six cluster algorithms are applied to separate the dataset into groups. Without providing true cluster of each record in the data set, which means pure unsupervised cluster, we cannot evaluate the performance by calculating the distance between predict cluster and true label. Four evaluation metric are given to evaluate the performance of unsupervised clustering algorithms. The tables given in last section are the results.

1. Silhouette Coefficient (SC)

The Silhouette Coefficient is defined for each sample and is composed of two scores:

a: The mean distance between a sample and all other points in the same class.

b: The mean distance between a sample and all other points in the next nearest cluster.

The Silhouette Coefficient s for a single sample is then given as:

The score is bounded between -1 for incorrect clustering and +1 for highly dense clustering. Scores around zero indicate overlapping clusters.

The score is higher when clusters are dense and well separated, which relates to a standard concept of a cluster.

2. Calinski-Harabasz Index (CHI)

The index is the ratio of the sum of between-clusters dispersion and of inter-cluster dispersion for all clusters (where dispersion is defined as the sum of distances squared)

The score is higher when clusters are dense and well separated, which relates to a standard concept of a cluster.

The score is fast to compute.

3. Davies-Bouldin Index (DBI)

This index signifies the average ‘similarity’ between clusters, where the similarity is a measure that compares the distance between clusters with the size of the clusters themselves.

Zero is the lowest possible score. Values closer to zero indicate a better partition.

The computation of Davies-Bouldin is simpler than that of Silhouette scores.

The index is computed only quantities and features inherent to the dataset.

The Discussion focuses on the research question. This section is where I will interpret the results in last section, account for the findings and explain their significance within the context of other research.

Consider the adequacy of the sampling techniques, the scope and longevity of the study, any problems with data collection or analysis and any assumptions on which the study was based.

**Conclusion**

This part will follow on naturally from key points raised in the discussion. The significance of the findings should be discussed in this section.

**Reference list**