Networks and Spatial Economics (2023) 23:373–406
https://doi.org/10.1007/s11067-021-09539-4

Analysing Spatial Intrapersonal Variability of Road Users Using Point-to-Point Sensor Data

F. Crawford [1] · D. P. Watling [2] · R. D. Connors [2,3]

Accepted: 14 April 2021 / Published online: 1 July 2021

Abstract

The availability of newly emerging forms of data in recent years has provided new opportunities to study spatial intrapersonal variability, namely the variability in an individual's destination and route choices from day to day. As well as providing insights into traveller needs, preferences and adaptive capacity, spatial intrapersonal variability can also inform the development of user classes for models of network disruption and for measuring behaviour change to evaluate the impact of network changes. This paper proposes a methodology for measuring spatial intrapersonal variability using point-to-point sensor data such as Bluetooth or number plate data. The method is innovative in accounting for sensor specific probabilities of detecting a passing device or vehicle and in providing a single measure for each traveller which considers destination and route choice variability and both the quantity of different trajectories utilised as well as the intensity with which they are used. A data science method is also presented for examining relationships between different trajectories observed in the network based on whether they are typically made by the same travellers. A case study using 12 months of real-world data is presented. The example provided demonstrates that a substantial amount of data processing is required, but the outputs of the methods are easily interpretable. Perhaps surprisingly, the analysis showed that the trips people made on weekdays were more evenly spread across a range of different trajectories than the trips they made during the weekend which were more concentrated into a few spatially similar clusters.

Keywords Intrapersonal variability · Association rule mining · Market basket analysis · Bluetooth data · Spatial variability

✉ F. Crawford
fiona.crawford@uwe.ac.uk

[1] Centre for Transport and Society, University of the West of England, Coldharbour Lane, Bristol, UK

[2] Institute for Transport Studies, University of Leeds, Leeds, UK

[3] Faculté des Sciences des Technologies et de Médecine, University of Luxembourg, Luxembourg, Luxembourg

1 Introduction

The availability of newly emerging forms of data in recent years has provided new opportunities to study spatial intrapersonal variability, namely the variability in an individual's behaviour from day to day in terms of destination and route choices. These types of insight into an individual's behaviour are important for three reasons. Firstly, by understanding spatial variability we gain insights into the needs, knowledge and behaviour of users. As well as providing background knowledge for decision makers, it can also be used in the development of transport model user classes based on attitude to risk (Shao et al. 2006) or information availability (Han et al. 2018), for example. Another application could be the design of transport policies involving spatial boundaries such as fare zones or congestion charging zones. Secondly, spatial intrapersonal variability provides insights into individuals' knowledge of the network which plays a role in determining traveller behaviour during transport disruptions (Papangelis et al. 2016, p.63). Measures of spatial regularity and/or variability could therefore be used to assess the adaptive capacity of network users (Wang 2015) and to inform models of traveller response to network disruptions. Thirdly, measures of spatial intrapersonal variability provide insights into the predictability of traveller behaviour and thus could inform the parameter values for day-to-day dynamical models which include learning mechanisms, for example the "switching choice probability" (Cantarella and Cascetta 1995) which relates to travellers reconsidering, but not necessarily changing, their previous route choice.

Previous research on spatial intrapersonal variability has mostly focused on the extent of travel or variability in the locations visited or routes taken. The extent of an individual's travel over a period of time can be represented by an activity space. Activity spaces can take different forms based on the method used to generate them which could include Daily Path Areas (Hirsch et al. 2016), confidence ellipses, a kernel density approach or shortest path networks (Schönfelder and Axhausen 2003). The size of individuals' activity spaces can then be compared or the characteristics of the activity spaces can be measured using built environment variables such as density of destinations, green space, land use or transportation facilities (van Heeswijck et al. 2015). The data used in this type of research often comes from travel diaries (Dijst 1999; Susilo and Kitamura 2005; Schönfelder and Axhausen 2003), although emerging forms of data such as mobile phone data (Järv et al. 2014) and GPS surveys (Hirsch et al. 2016) have also been used.

Activity spaces do not generally consider the intensity of travel within the space, although a few researchers such as Järv et al. (2014) have created separate activity spaces for daily and less frequent activities. Accounting for the intensity of travel within an activity space is essential for understanding intrapersonal variability, as it provides insights into how well the traveller knows that part of the network and the predictability of their travel from day to day.

A separate branch of the literature on spatial intrapersonal variability focuses on the origins and destinations of trips only and uses a more statistical than geographical approach. Often this type of analysis has been applied to travel diary data (Muthyalagari et al. 2001; Schlich et al. 2004) or smartcard data (Kieu et al. 2015;

Goulet-Langlois et al. 2016). The focus on locations visited has also been used with other types of data to simplify the spatial data, such as number plate data (Chen et al. 2017) and mobile phone data (Masso et al. 2019). Measures of variability include distance travelled (Muthyalagari et al. 2001), number of locations visited (Masso et al. 2019), number of different OD pairs (Chen et al. 2017), share of destinations within a given distance from home and number of 'new' destinations visited per day (Schlich et al. 2004). Buliung et al. (2008) define a spatial repetition index which divides the number of activities undertaken at repeated locations by the total number of activities undertaken by an individual traveller.

Intrapersonal variability in route choice has typically been undertaken on smartcard data (Kurauchi et al. 2014; Kim et al. 2017), although some small GPS surveys have also been undertaken to examine the degree of habit in route choices of road users (Vacca et al. 2019). Route choice variability is usually calculated separately for each OD pair, which is useful for understanding service choices through a public transport network at an aggregate level but it does not generate a single measure for each traveller which could provide insights for a road network manager, for example.

To understand the entirety of a traveller's spatial intrapersonal variability, it is necessary to consider both the locations visited and the routes taken. Only a few papers have considered both aspects. Shen et al. (2013) considered both location and route choice intrapersonal variability in commuting trips using a seven day GPS survey. They used a binary representation of these two types of variability and also for temporal and modal variability to produce seven commuter types. Crawford et al. (2018) proposed a methodology for identifying road user types based on trip frequency, spatial and temporal variability. The spatial aspect was measured by identifying clusters of 'spatially similar' trips based on detections at fixed Bluetooth sensors. Two measures of spatial intrapersonal variability were used: the number of clusters used by a traveller in the given period of time and the percentage of a traveller's trips in their most commonly used cluster.

For all of these approaches, newly emerging sources of data are providing new opportunities to gain better insights into intrapersonal variability. The current research is timely due to two separate impacts of developments in Information and Communication Technologies (ICT). The first impact is the emergence of new passively collected data sources relating to mobility, including the mobile phone and Bluetooth data mentioned above. Studies such as Järv et al. (2014) which used a whole year of data for a large sample of people from a city would not have been feasible previously. The second impact of ICT has been on travel itself. More people can now work, at least occasionally, remotely (Felstead 2012) and freely accessible real-time information plays a role in route choice. ICT is also enabling new mobility services, many of which operate on an on-demand basis. For such services to be economically viable, it is crucial that everything from the charging structures to the organisation of vehicles and staff are designed based on traveller needs, which includes their multiday behaviour.

The current research focuses specifically on road users as very little of the previous research has focused on this group, despite the large number of travellers and trips involved. To provide useful insights into traveller behaviour and network knowledge for network operators it is necessary to examine variability in both ODs and routes. Previous research typically includes a number of measures of spatial variability but for ease of interpretation it would be preferable to have a single measure which takes into

account the range of trips made as well as the intensity with which different trips are made. The existing literature can be separated into area based and user based analyses and therefore there is also a gap in terms of a 'trajectory' focus (to use the terminology of Toch et al. (2019)). This research, therefore, will also propose a method to examine multi-day trip behaviour by measuring the relationships between different possible trip trajectories based on the overlap in people making those types of trip.

Point-to-point sensor data will be used for this research as it provides detailed spatial information and many road network managers have access to such data. Types of point-to-point data which are already widely collected along the road network include Automatic Number Plate Recognition (ANPR), Bluetooth, and Electronic Toll Collection (ETC) data. This choice of data will be discussed further in Section 2. A suitable method for processing point-to-point data and assigning every trip (a sequence of observations) to a cluster of spatially similar trips has been presented in Crawford et al. (2018). The current paper utilises the processing steps proposed, but extends the clustering process to account for detectors which have different probabilities of detecting a passing device or vehicle. Whilst Crawford et al. (2018) used two separate measures to represent spatial intrapersonal variability, the current paper takes inspiration from the field of ecology to propose a single measure which combines the number of different trajectories made through the network and the distribution of trips between those trajectories. Unlike the previous research, the current research also moves beyond the user perspective by using a technique from data science to provide insights into repeated trip making in a way which is directly connected to the road network.

The contribution of this paper is to propose a methodology which can be used by road network managers to gain insights into spatial intrapersonal variability from a traveller perspective and a network perspective. A single measure is proposed to represent OD and route choice variability. Association rule mining will also be used to examine intrapersonal variability in transport for the first time, to the authors' knowledge.

The structure of the paper will be as follows. Section 2 discusses point-to-point sensor data in more detail and also raises some issues in relation to Bluetooth data, which will be used in the case study section. Section 3 outlines existing methods which form the basis of the proposed methodology. Section 4 presents the methodology proposed by the authors as a means of gaining new insights into spatial intrapersonal variability. Section 5 applies the methodology to a real-world case study in northern England. Section 6 discusses possible applications. Finally, Section 7 describes future research directions and concludes the paper.

2 Data for Examining Spatial Intrapersonal Variability

Different data sources are more suitable for providing information about particular kinds of intrapersonal variability. Travel diaries can be useful for examining variations in ODs as the purpose of travel can also be collected (Huff and Hanson 1986; Schlich et al. 2004; Bayarma et al. 2007). Newly emerging forms of data typically do not rely on participant recall and can be collected for longer periods of time and with lower costs and participant burden than self-completion travel diaries. In the literature, many travel diaries collect data for 15 days or less (Jones and Clarke 1988; Stopher and

Zhang 2011; Safi et al. 2015) whilst other diaries cover longer periods of time (between 35 days and 1 year), but have relatively few participants (149 people and 139 and 153 households respectively for Huff and Hanson (1986), Bayarma et al. (2007) and Elango et al. (2007)). In contrast, passively collected data can be analysed for periods of a year or more as shown in Järv et al. (2014) and Crawford et al. (2018).

Telecom companies passively collect data from mobile phones which can be used for travel behaviour research. Different kinds of data are available including event-driven data, such as Call Detail Records, and network-driven data, which often has higher spatial resolution (Wang et al. 2018, p.143) but is far less accessible to researchers. Mobile phone data is therefore useful for measuring dynamic population densities (Calabrese et al. 2011) and aggregate level OD matrices for fairly large cell tower areas. At an individual level, it can be used to analyse individuals' activity spaces over twelve month periods or longer (Järv et al. 2014). A major challenge with data of this type is accessibility due to data protection regulations and also the costs in obtaining data from some providers.

Another valuable form of spatial data comes from Global Positioning System (GPS) devices. GPS data can provide reliable information with good geographical precision, particularly when effective data processing is undertaken (Schuessler and Axhausen 2009). Whilst GPS data has been used in travel behaviour research, it typically takes one of two forms. Firstly, many studies have been undertaken which provide a group of participants with GPS devices to either keep on their person or in their vehicle for the study period. Many of these studies involved some sort of diary alongside the GPS data capture and they mostly cover a period of one week or less (for example Ramaekers et al. (2013), Houston et al. (2014) and Millward et al. (2019)). Secondly, other studies have utilised passively collected GPS data relating to vehicles used for passenger transport, particularly buses and taxis (Liu et al. 2015; Shen et al. 2018; Tu et al. 2018). Such data is useful, particularly in terms of travel times, but they only provide information about a non-representative set of trips and typically only vehicles can be matched across days, not passengers. Data protection regulations mean that GPS data is very difficult to obtain for non-commercial vehicles or personal devices, unless devices are provided to participants for that purpose.

Stock (2018) provides a summary of the use of social media data for examining locations. Their detachment from transport networks and lack of spatial precision mean that mobile phone data and social media data can provide better information about activity spaces, zone attractiveness (Moya-Gómez et al. 2018) and land use (Zhan et al. 2014) than detailed information about spatial intrapersonal variability.

Data is also collected routinely in the operation of transport systems, particularly when users pay for access. A large proportion of the literature examining spatial intrapersonal variability using empirical data is focused on public transport smartcards (Kurauchi et al. 2014; Kieu et al. 2015; Kim et al. 2017). The spatial data availability from such systems varies, with some systems providing boarding and alighting data and others providing only one of the two (Kurauchi et al. 2014, p.24). Systems such as the London Underground require travellers to tap in and to tap out of the network, but they cannot provide data about the exact service or route used, unlike bus services (Kurauchi et al. 2014, p.26).

This research focuses on point-to-point data, where fixed sensors store unique identifiers of passing people or vehicles so that observations from different sites can

be connected. Types of point-to-point data include Automatic Number Plate Recognition (ANPR), Bluetooth and Electronic Toll Collection (ETC) data. Future examples could include data from vehicle to infrastructure communications. Point-to-point sensor data has the advantage that large quantities of data can and are being collected in this form around the world and perhaps more importantly, they are being collected by transportation practitioners directly. This means that we generally have greater levels of control over the data collection and the processing undertaken than for other sources of data. Point-to-point sensor data also has good geographic precision with a close network connection and the fixed nature of the sensors means that comparable data can be collected for different users and/or different days or even months. As with many of the emerging sources of data, this is a passively collected source of data which has minimal burden on the public and does not rely on participant recall.

2.1 Point-to-Point Sensor Data

Point-to-point data has been used by researchers for many years in the form of manually collected number plate surveys (Watling et al. 2012), but technological advances mean that data can now be collected in bulk and new forms of point-to-point sensor data have emerged. Automated Number Plate Recognition (ANPR) systems are now available which can 'read' number plates from roadside cameras and therefore data can be collected over longer periods of time and at more locations. ANPR data has been used to measure trip frequency, time of day intrapersonal variability and OD variability (Chen et al. 2017; McLeod et al. 2017). Bluetooth and WiFi capabilities are now in a multitude of personal device and in-vehicle systems. By setting up detectors to constantly scan for nearby devices, Bluetooth-enabled and/or WiFi-enabled devices can be identified and data on movements through space can be recorded (Versichele et al. 2012; Traunmueller et al. 2018). Electronic Toll Collection systems can also provide insights into when and where users travel (Kim et al. 2014; Tam and Lam 2008).

There are also similarities with smartcard data, although in that case the detection points are at the entry and exit points to an underground network, for example, or the boarding points for bus services. The detection locations, therefore, are not determined by data collection requirements alone. Also, the number of detections made during a trip are much lower and could include just one observation (for example boarding a bus), two observations (entering and exiting a subway system) or more if the traveller interchanges many times. The type of methods described above relating to OD data are generally better suited to this sort of data than the methodology presented below, unless lots of interchange data is available.

As discussed in Crawford et al. (2018), despite the advantages there are also challenges in using point-to-point sensor data for examining spatial intrapersonal variability. Point-to-point sensors typically only record the unique identifier associated with a passing traveller, vehicle or device together with the timestamp of the detection. This type of data, therefore, does not include trip purpose information. This information is not necessary for the uses outlined for this research, including the analysis of familiar routes and the measurement of the impact of interventions. It does mean, however, that personal travel cannot be separated from business travel, which may mean that the results are not directly comparable with analyses using travel diary data, which in

England typically only records personal travel. By undertaking research containing all motorised vehicles on the road, including buses and taxis, the current analysis could provide valuable insights for road network managers.

Also, the data does not include origin or destination information, either in terms of the location or the time. We cannot, therefore, measure or even estimate an individual's departure time. What the data can tell us, however, is the regularity at which a device passes a particular location. The reason for any variability could be differences in departure time and/or differences in traffic conditions encountered prior to passing the detector. Despite initially appearing problematic, this may be beneficial from the viewpoint of local road network managers. Their priorities lie with road users at the point that they enter their jurisdiction and on critical links within the network. The road managers can, therefore, design the Bluetooth detector placement to focus on the relevant parts of trips only. The use of fixed detectors also means that analysts can examine the variability in the times of day at which a road user passes a fixed point, for example a pinch point such as a bridge or tunnel or a location related to an intervention such as a charging cordon.

2.2 Bluetooth Data

This paper will use Bluetooth data in the Case Study presented in Section 5. The data is from 2015 and therefore more recent problems with re-identifiability of devices due to technological developments for privacy reasons are not relevant. The proposed methodology is relevant for all types of point-to-point sensor data.

Fixed Bluetooth sensors, also known as detectors, can be placed alongside roads and then set to continuously scan for any discoverable Bluetooth devices within their detection zone (see Bhaskar and Chung (2013) for more details). They record the unique identifier (known as a MAC address) and corresponding timestamp for discoverable Bluetooth devices passing close by. The Bluetooth devices could be associated with the vehicle, for example in-car sound systems or hands-free kits, or with a person in the vehicle, for example a mobile phone, tablet or laptop. Sensors are installed at fixed locations and therefore the geographic coverage is defined a priori by the sensor locations, rather than being determined by the traveller as is the case with GPS or mobile app data. As the data is passively collected, there are disadvantages, however, as additional data cannot be requested from participants, such as trip purpose or mode of travel.

Data from fixed Bluetooth detectors is becoming increasingly popular for measuring travel times on the road network (Haseman et al. 2010; Hainen et al. 2011; Moghaddam and Hellinga 2013), particularly in urban areas, and has also been used in OD estimation (Barcelö et al. 2010; Carpenter et al. 2012). Spatial data can also be collected for individuals over multiple days using Bluetooth sensors (Delafontaine et al. 2012; Crawford et al. 2018; Traunmueller et al. 2018) as the sensors record device specific identifiers. Issues relating to Bluetooth data are relatively well known in terms of their impact on estimating travel times, but their impact on estimates of repeated trip behaviour has not been examined.

Observations from Bluetooth detectors relate to *devices* and not to people or vehicles. For travel time estimation this is taken into account by appropriate data cleaning to retain only motor vehicle movements and by examining the effects of

having multiple devices within a single vehicle (for example in Bhaskar et al. (2015)). For examinations of intrapersonal variability, the fact that it is devices being recorded needs to be considered when interpreting results.

While in some cases devices are likely to remain close to one person, for example mobile phones, other types of device may be shared by households or businesses, for example in-car systems. For repeated trip analyses, having individual level data could be considered preferable as it would be comparable with personal travel diary data. In some cases vehicle level data may be more informative, however, for example by recording what Zhang et al. (2002) call 'allocated' household activities, whoever undertakes them. In other cases vehicle data may be misleading, for example the widely reported use of Bluetooth within rental cars (for example USA Today (2015)). Millard et al. (2016) found that only 16% of car rentals are for more than a week and therefore Bluetooth devices in these cars are likely to demonstrate more variability over longer periods of time as multiple customers would be associated with the same unique MAC address. Rental cars are, however, likely to make up a very small percentage of the Bluetooth sample in most cases.

When estimating travel times, if the Bluetooth penetration rate is sufficiently large, any bias in the travellers with Bluetooth devices is assumed to have a minimal impact on the estimates. When examining intrapersonal variability, however, a biased sample of travellers may have a significant impact on results as aggregation does not occur prior to analysis. People with personal Bluetooth devices or vehicles with Bluetooth-enabled features may have higher incomes and lower ages than those who don't, which has implications for the representativeness of the data from this source. The number of car trips per year (as a driver) increases between the ages of 17 and 49 and then falls with age, according to Department for Transport (2016a), so Bluetooth data may not provide a representative sample of trip frequency. Minnen et al. (2015) found differences in day-to-day variability in travel behaviour by age, where people aged 25–45 had lower levels of variability possibly due to a higher number of constraints on their time, which means that Bluetooth data may also be biased in terms of intrapersonal variability. Socio-economic classification also has an impact on the number of trips made by car and the distance travelled (Department for Transport, 2016b). Elango et al. (2007) found that higher income households have greater variability in travel behaviour. In contrast, however, Minnen et al. (2015) found more variability in travel patterns for unemployed people, compared to employed people, perhaps driven by the differences in trip types made, although only five days of data were analysed in that research. In summary, if Bluetooth-enabled devices are more prevalent amongst younger people and people with higher incomes, then the observed travel patterns may be biased in terms of frequency and variability and this should be taken into account when designing policies based on the analysis.

Fixed Bluetooth sensors will not detect every discoverable Bluetooth device which passes through its detection zone. The probability that a discoverable Bluetooth device will be detected depends on many factors, some of which vary between locations and some of which vary over time. Detection probabilities can depend upon the speed at which the Bluetooth device is travelling, the device manufacturer, the number of other devices in the detection zone, physical barriers, the weather, the location and position of the sensor (such as the height (Brennan et al., 2010)), and the set-up of the sensor (Araghi et al., 2014, Michau et al., 2014, Michau et al., 2017, Tsubota and Yoshii,

2017). Sensor specific detection probabilities have not been taken into account in previous research looking at intrapersonal variability using Bluetooth data, for example Crawford et al. (2018).

3 Method Selection

The aim of this paper is to propose a methodology which can be used by road network managers to gain insights into spatial intrapersonal variability from a traveller perspective and a network perspective using point-to-point sensor data. There are a number of different effects which point-to-point sensor data could be used to examine in terms of spatial variability. One could consider changes over time, within person variability or relationships between different parts of the network, for example. The current research considers the latter two points. In the data analysed, we assume that there is no systematic change in travel behaviour over time and therefore the order in which trips are observed is not taken into account. Methods which treat the data as panel data could provide useful insights, but that is not the focus of the current paper.

This section will consider the qualities required in the methods selected and the different options available. As it should be practical to apply the methodology to large amounts of Bluetooth data, three types of methods are required:

1. Dimensionality reduction,
2. Calculation of spatial intrapersonal variability from a traveller perspective and
3. Measurement of spatial relationships in the network.

These three processes will now be considered in turn.

3.1 Dimension Reduction

Once point-to-point sensor data has been cleaned, individual observations can be connected into trip trajectories if the same device is detected at multiple sensors and the time between observations is consistent with driving between the two locations. A balance needs to be struck when examining the data between the amount of spatial information which is retained relative to the complexity and usefulness of outputs. This is particularly true when analysing very large sets of data for a town or for several months or years, as each trip in the data will have been observed at *at least* two sensor locations. The aim of this process is to reduce the dimensionality of the spatial aspect of the data.

This research focuses on the concept of 'trips' within the Bluetooth data, where a trip is a sequence of observations of the same Bluetooth-enabled device as it moves within the network. This provides greater spatial insights than looking at each site where a particular Bluetooth device was detected independently. Sensors or cameras collecting point-to-point data are also typically installed for reasons other than examining spatial variability, for example Bluetooth detectors are used to measure travel times and ANPR cameras are used for speed limit enforcement and cordon charging. Therefore, whilst the data is collected on key routes, it is often not collected at 'interesting' locations in terms of predicting the destination activity.

Dimension reduction could be undertaken by using the first and last sensor where the device was detected for each trip. This would, however, result in the loss of information about route choice. With Bluetooth data there is also the issue that a non-trivial percentage of Bluetooth-enabled devices passing a sensor will not be detected. This makes the first-last sensor approach less appealing since these are not necessarily the first and last Bluetooth sensors that were passed on that trip.

A natural solution would be to cluster together similar trips. By using unsupervised learning, groupings may arise which would not have been identified by applying rules simply based on Traffic Analysis Zones, for example. The challenge is to determine what measure of similarity/distance to use to compare two trip trajectories.

One method which has been used successfully for this purpose is Sequence Alignment. This method comes from Bioinformatics but it has also been used in the social sciences (Abbott 1995; Shoval and Isaacson 2007). In transportation research, the method has been used for the comparison of daily or weekly activity patterns or time use data (Wilson 1998; Joh et al. 2001, 2002; Dharmowijoyo et al. 2017), though that application is quite different to the trip trajectory comparison considered in the current research. For trip trajectories, where reordering is of less importance, the more relevant literature compares movements using point-to-point sensor data (Delafontaine et al. 2012; Versichele et al. 2012; Crawford et al. 2018) or vehicle traces (Kim and Mahmassani 2015).

Sequence Alignment takes two sequences of letters or 'strings' as an input and after identifying the optimal alignment between the two strings (which may involve inserting gaps known as indels into one or both strings), the 'distance' or cost associated with the optimal alignment is outputted. In the current application, the sequences are made up of letters, each of which represents a Bluetooth detection at the site assigned with that letter.

There are two types of Sequence Alignment techniques - global techniques, which attempt to match entire sequences, and local techniques, which look for parts of the two sequences which match. As in Crawford et al. (2018), global alignment will be used in the current paper as we seek to identify trajectories which are similar in their entirety, rather than looking for partially overlapping trajectories as in Kim and Mahmassani (2015). As shown in Fig. 1, this requires the alignment of each letter in both sequences with either a letter in the other sequence, or an indel. The optimal alignment minimises the total cost, which is the sum of each of the pairwise costs. The cost associated with aligning two letters is defined as the on-road distance between the two sensor locations.

The methodology allows trip sequences consisting of differing numbers of observations to be compared. This is made possible by assigning a special distance or cost to the alignment between a letter in one sequence and a gap (called an 'indel') in the other sequence. This can occur within a sequence (as in Fig. 1) or at the start or end of a sequence.

Crawford et al. (2018) used a fixed indel cost, irrespective of the Bluetooth sensor it is aligned with. They used a cost equal to half of the distance between the two furthest

Sequence 1:	A	B	C	-	E	F
Sequence 2:	A	B	C	D	E	F

Fig. 1 Example trip sequences

apart sensors in the network. This was considered to be the optimal value as a smaller indel cost would result in the distance between distant sensors not been fully accounted for as each of the sensors would instead be matched with an indel as shown in the lower diagram in Fig. 2. A larger indel cost is not used as it would result in sequences being disproportionately clustered based on their length rather than their contents.

Once sequence alignment has been used to determine the distance matrix for all trip sequences in the data, hierarchical clustering can be undertaken to identify the groups of spatially similar trips.

3.2 Spatial Intrapersonal Variability (Traveller Perspective)

As discussed in Section 1, spatial intrapersonal variability has been examined in a number of different ways. The requirements for the current research are that a single measure can be produced for each traveller, the measure should take into account the number of different categories of trips undertaken (in terms of their spatial qualities only), and also the distribution of the traveller's trips between the different categories of trips.

The requirements are similar to those for analysing multimodal behaviour, although in this case the categories are the groups of spatially similar trips rather than modes of transport. Diana and Pirra (2016) discuss a number of suitable methods from different fields, including measures of entropy, inequality and species diversity. Whilst a number of these methods may prove suitable for our purposes, this research uses the Herfindahl-Hirschman Index (HHI). The HHI is often used for examining the market share of different businesses. It is also known as Simpson's Diversity Index in ecology, where it is used to produce measures of species diversity. This measure was selected for the current paper as it focuses on 'dominance', as the market shares (or the proportions of trips in each spatial cluster in our case) are squared. The HHI has also been used for other purposes within transportation research, including for mode choice (Heinen and Chatterjee 2015; Susilo and Axhausen 2014) and public transport route choice (Kim et al. 2017).

The normalised Herfindahl-Hirschman Index can be calculated as follows:

$$H^* = \frac{\left(\sum_{i=1}^{M} s_i^2\right) - \left({}^1/_M\right)}{1 - \left({}^1/_M\right)} \tag{1}$$

where M is the total number of businesses, and s_i is the market share of business i.

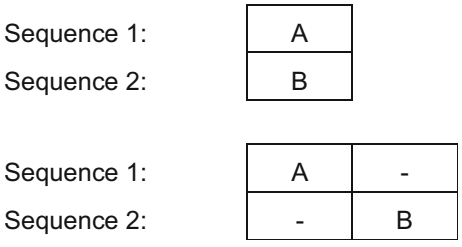

Fig. 2 Two possible alignments of observations at sites A and B

3.3 Spatial Relationships (Trajectory Perspective)

The Herfindahl-Hirschman Index provides a measure of the spatial intrapersonal variability for travellers but alongside this it would be valuable to explore the relationships between the different spatial clusters based on whether they are used by the same people. This could identify segregation within the network, with different parts of the network being exclusively used by different groups of people and it could also be used to examine the extent to which two road network interventions in a city benefit the same group of people.

By defining activity spaces for each traveller, as in Järv et al. (2014), it would be possible to visualise and quantify overlaps in travellers' activity spaces. As with the discussion of activity spaces in Section 3.1, however, this would be limited in that it ignores the direction of travel and it may miss more subtle differences in routes or areas visited. Instead, this subsection proposes a method known as association rule mining (Agrawal et al. 1993; Tan et al. 2014, Chapters 6 and 7). This approach is also known as Market Basket Analysis as it has traditionally been used to provide insights into the products which are commonly purchased together in shops, for example bread and butter. It has been used in transportation research to examine relationships between road traffic crash characteristics (Pande and Abdel-Aty 2009) and to explore which shops are typically visited by the same customer whilst in the city centre (Yoshimura et al. 2018). To the authors' knowledge, it has not previously been used to examine repeated trip behaviour.

Association rule mining examines which *items* frequently appear in the same *transaction*, i.e. people 'purchase' these items together. Each transaction is a set of items purchased, with each item represented at most once. The aim is to identify association rules of the form $X \rightarrow Y$, where people who buy the antecedent item set (X) are also likely to buy the consequent item set (Y). For example, one might find the following rule in data from an electronics store:

$$\{printer\} \rightarrow \{printer\ cartridge, printer\ paper\}$$

Here, people buying a printer typically also buy ink and paper for their new printer. It is a directional relationship as people buying ink and paper together are often not also buying a printer.

The antecedent and the consequent in an association rule are item sets containing one or more item. The same item cannot appear in both the antecedent and the consequent. Whilst rules can be written for all possible item sets, it is only the rules which demonstrate a strong association between X and Y which are informative. The three measures shown in eqs. (2) to (4) can be used to assess the strength of the association rule $X \rightarrow Y$. Support measures the proportion of transactions containing both X and Y, thus helping us to avoid rare cases. Confidence indicates how often Y is true when X is true. Lift measures the strength of the relationship between X and Y.

$$Support = \frac{\sigma(X\ and\ Y)}{N} \tag{2}$$

$$Confidence : \delta(X \rightarrow Y) = \frac{\sigma(X \text{ and } Y)}{\sigma(X)} \tag{3}$$

$$Lift = \frac{\delta(X \rightarrow Y)}{\sigma(Y)} \tag{4}$$

where $\sigma(X)$ is the count of all transactions containing item set X, and N is the total number of transactions analysed.

Whilst thresholds for support and confidence can be used to identify meaningful rules, it would be impractical to calculate these measures for all possible association rules for a large dataset. If we only had four possible items (1,2, 3 and 4), then there would be 50 possible association rules (see Table 1) and the number of possible rules increases exponentially with the number of items.

The Apriori Algorithm provides an efficient method for identifying association rules with support above a given threshold (Agrawal and Srikant 1994). The algorithm 'prunes' the item sets used to create rules by identifying in advance which will have insufficient support. The properties which are utilised are:

1. If an item set I has sufficient support, then every subset of I also has sufficient support, and
2. If an item set I has insufficient support, then all of its supersets will also have insufficient support.

The process begins by considering each item separately and calculating the support for each one. Any items with insufficient support should then be removed. Item sets containing two items would then be created using only the items with sufficient support in the previous stage. The support for each of these item sets would then be compared

Table 1 All possible association rules when there are 4 items

{1} → {2}	{1} → {2,3}	{1} → {2,3,4}	{2,4} → {1}	{1,3,4} → {2}
{1} → {3}	{1} → {2,4}	{2} → {1,3,4}	{2,4} → {3}	{2,3,4} → {1}
{1} → {4}	{1} → {3,4}	{3} → {1,2,4}	{3,4} → {1}	
{2} → {1}	{2} → {1,3}	{4} → {1,2,3}	{3,4} → {2}	
{2} → {3}	{2} → {1,4}	{1,2} → {3}	{1,2} → {3,4}	
{2} → {4}	{2} → {3,4}	{1,2} → {4}	{1,3} → {2,4}	
{3} → {1}	{3} → {1,2}	{1,3} → {2}	{1,4} → {2,3}	
{3} → {2}	{3} → {1,4}	{1,3} → {4}	{2,3} → {1,4}	
{3} → {4}	{3} → {2,4}	{1,4} → {2}	{2,4} → {1,3}	
{4} → {1}	{4} → {1,2}	{1,4} → {3}	{3,4} → {1,2}	
{4} → {2}	{4} → {1,3}	{2.3} → {1}	{1,2,3} → {4}	
{4} → {3}	{4} → {2,3}	{2,3} → {4}	{1,2,4} → {3}	

against the threshold and any with insufficient support would be removed. The process would continue until there are no further item sets to combine.

This is more efficient then calculating the support for all item sets. If in the example above, the set {1} had insufficient support, then all of its supersets could immediately be discarded, including {1,2}, {1,3}, {1,2,3} etc..

The manner in which these methods can be used to examine spatial intrapersonal variability will be discussed in the following section.

4 Methodology

To gain insights into spatial intrapersonal variability from point-to-point sensor data, this paper proposes the methodology presented in Fig. 3.

This section will focus on the three innovative stages in the methodology, namely the process of identifying clusters of spatially similar trips (Section 4.1), measuring spatial intrapersonal variability for each traveller (Section 4.2), and exploring relationships between different trajectories within the network based on the people who are using them (Section 4.3).

4.1 Dimension Reduction

The dimensionality of the data can be reduced by clustering spatially similar trips as discussed in Section 3.1. The methodology proposed in Crawford et al. (2018) takes into account the non-zero probability of a Bluetooth-enabled device not being detected as it passes a sensor but it does not take into account different sensors having higher or lower probabilities of detecting a passing Bluetooth-enabled device. This can occur

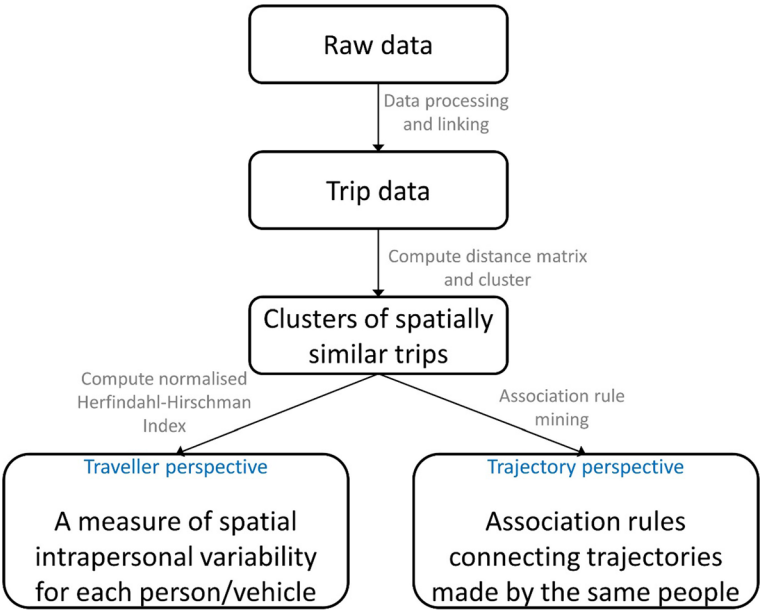

Fig. 3 Overview of proposed methodology for analysing spatial intrapersonal variability

with Bluetooth detectors due to variations in the height of installation, distance from the centre of the road and nearby infrastructure. The current paper therefore extends the methodology by introducing sensor specific indel costs.

An indel cost represents the distance when aligning an observation in one sequence (for example at Site A) against a gap inserted into the other sequence. A sensible adjustment would be to lower the indel cost when aligning a sensor with a lower detection rate against a gap, since there is a higher likelihood that the device did pass the sensor but was not observed therefore we do not want to excessively penalise a missed observation.

For example, consider the two link network in Fig. 4. A, B, C and D are Bluetooth detectors. All vehicles contain one Bluetooth-enabled device and they travel from A to D. Let us assume that 80% of Bluetooth-enabled devices are detected when passing B but only 50% of Bluetooth-enabled devices are detected when passing C. A and D detect all Bluetooth-enabled devices. If 100 vehicles make the trip ABD and 100 vehicles make the trip ACD, then we would expect approximately 130 observations at either B or C. We would record 80 ABD trips, 50 ACD trips, and 70 AD trips. Sequence alignment is used in this paper to determine the distance matrix for the hierarchical clustering of trip sequences. By applying a constant indel cost in the sequence alignment process, we implicitly assume that the 'distance' between trip sequences AD and ABD and the 'distance' between AD and ACD are the same. We know, however, that this is not the case and that the trips only recorded at AD are more likely to represent ACD trips, and so the distance between AD and ACD should be smaller. In this case, the sensor specific indel cost for aligning with sensor C should be 0.625 (50%/80%) times the indel cost for aligning with sensor B.

The current research maintains the upper bound on indel costs proposed in Crawford et al. (2018), namely half of the shortest path between the two furthest apart sensors. Sensor specific indel costs can then be calculated relative to this upper bound, with the sensor with the highest detection rate having an indel cost equal to the upper bound. The indel cost for sensor A can therefore be calculated using eq. (5).

$$IC_A = \frac{P(A)}{max_i(P(i))} \times \frac{max_{j,k}(dist(j,k))}{2} \tag{5}$$

Where $P(i)$ is the probability of a Bluetooth-enabled device being detected at sensor i and $dist(j,k)$ is the shortest distance by road between sensors j and k.

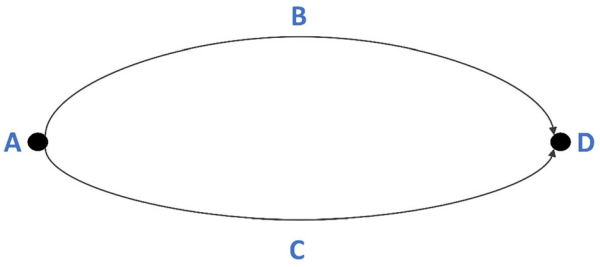

Fig. 4 Two link example

Whilst the calculation of the sensor specific indel costs is very straightforward, estimating the probability of detecting a Bluetooth device at each sensor is more challenging. In the case study below, sets of three Bluetooth detectors on one link are used to estimate the relationship between the number of detections per lane and the detection rate. Other options include undertaking experiments to directly measure the detection rates at each sensor location or utilising co-located Bluetooth sensors and loop detectors to estimate each sensor's detection rate.

4.2 Spatial Intrapersonal Variability (Traveller Perspective)

The spatial clusters create a more manageable number of categories for calculating the Herfindahl-Hirschman Index (HHI). The HHI can be used to measure the diversity of trips made by each traveller. The normalised HHI shown in (1) can be calculated separately for each traveller, using the total number of spatial clusters (M) and the proportion of the traveller's trips in each cluster (s_i).

The normalised version of the index is used in this research as it results in a value between zero and one, which makes interpretation easier. A normalised HHI of zero for a traveller represents an equal number of trips in all spatial clusters and a value of one represents all trips in one spatial cluster.

As well as calculating the HHI using all trips made by the traveller to produce a single measure of spatial diversity, the HHI can also be applied to subsets of trips made by each traveller to allow intrapersonal comparisons. In this paper, the spatial diversity of trips made on weekdays is compared against the spatial diversity of trips made on weekend days for each traveller separately. The same approach is used to compare the spatial variability in trips made in the summer of 2015 against those made in the autumn of 2015, again for each traveller separately.

4.3 Spatial Relationships (Trajectory Perspective)

After undertaking the spatial clustering, association rule mining can be used to explore which trajectories through the network are typically made by the same travellers. The 'items' are trajectories made by the traveller, as represented by the spatial clusters. By using the spatial clusters, rather than the full trip sequences, we obtain a more manageable number of items for examination. Each 'transaction' (or 'basket') relates to the trips made by one traveller over the period of the study. The transaction consists of a list of the spatial clusters used by that traveller, with each cluster appearing at most once.

As each item, or spatial cluster in our case, can appear at most once in a transaction, this method does not distinguish between frequent and less frequent trips. In the current paper, therefore, only spatial clusters used on a regular basis were included in the transactions. Regular was defined as at least once per month on average.

The method also only considers the spatial aspects of trips; the timing of trips, in terms of time of day, time of year or the order of trips, is ignored.

4.4 Applying the Methods

The current research utilised the open source statistical software R (R Core Team 2019) for all of the data processing and analysis. A small, anonymised dataset and code to

perform the three methods described above are available here: https://github.com/ficrawford/Spatial-intrapersonal-variability-using-Bluetooth-data. It should be noted that data cleaning and processing is far more labour intensive and has a much longer run time than the analysis itself.

The data structure required in R for the analysis is shown in Fig. 5. All of the methods can be applied using other tools but different data structures may be required. Point-to-point sensor data is typically obtained in separate files for each sensor or camera – these are denoted by Site A, B and C below. Each observation will include a unique identifier, denoted by 'MAC' here, and a date-time stamp. The following data processing steps are then required:

1. Basic data cleaning and processing. This will depend on the type of data obtained, but for Bluetooth data it will typically involve retaining only one observation when a device passes the sensor, then matching up observations across sites to get trip sequences. Trip sequences are the set of sensors at which the device was detected whilst making one trip. The timestamps are crucial for this step as they enable the splitting of different trips within a day where the time between observations is outside of the expected range for travel by car or van between the two sensor locations at that time. In preparation for the sequence alignment process, each trip should be one row in the dataframe and the sensor names should be listed, using a new variable for each observation. This is not an efficient way to store the data, but it is required for the upcoming steps.
2. The unique trip sequences observed should then be collated, along with the number of times each trip sequence was observed in the data (all travellers combined). Trip sequences observed very few times can be removed from the analysis at this stage to reduce computation times.
3. Weighted cluster analysis is then undertaken using Ward's method and sequence alignment is used to compute the distance matrix (TraMineR package (Gabadinho

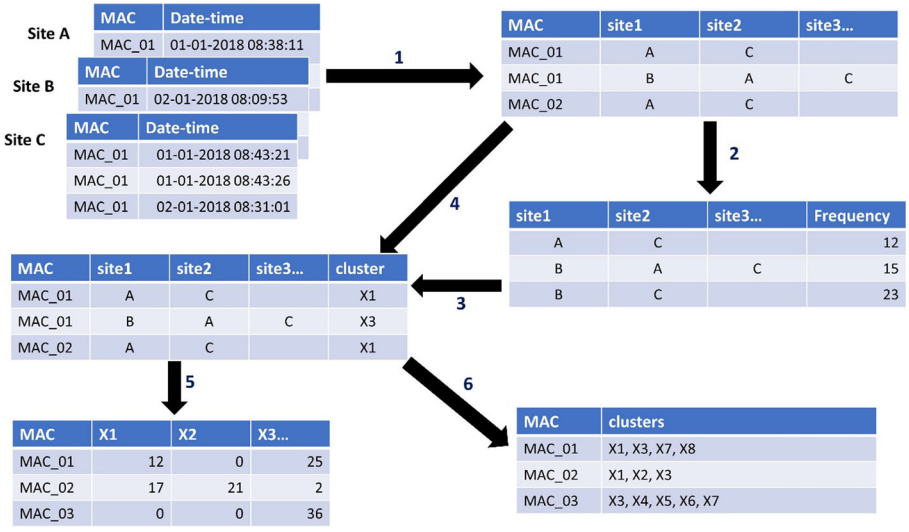

Fig. 5 Data structure used in R

et al. 2011a)). The as.clustrange function within the WeightedCluster package (Studer 2013) can be used to calculate measures to inform the choice of number of clusters.

4. The trip sequences and their assigned clusters should then be merged back into the full dataset so that the unique traveller identifiers can be used. The clustering did not take into account which trips were made by which travellers.
5. Before the HHI calculation can be made, the data must be processed so that each traveller has one row. For each traveller, the dataframe must record the number of trips made in each of the spatial clusters.
6. Before the association rule mining can be undertaken, the data must also be processed so that each traveller has one row. In this case, however, the data should be transformed into a list which includes the names of the clusters used frequently by that traveller (with each cluster name appearing at most once).

4.5 Factors Affecting Computational Complexity

As well as the data processing steps discussed in the previous subsection, there are three main steps in the methodology which require relatively large amounts of processing power for large datasets.

Firstly, there is the Sequence Alignment process. As this is used to populate the distance matrix for clustering spatially similar trips, Sequence Alignment must be undertaken for every pair of trip sequences. The complexity will therefore increase with the number of different trip sequences in the data and the pairwise comparisons will require more time and memory as the average number of Bluetooth sensors passed during a trip increases. The TraMineR package used in the current paper can undertake pairwise comparisons on 4318 sequences of average length 16 in 15 s (Gabadinho et al. 2011b, p.25).

Secondly, there is the hierarchical clustering process. This type of clustering is required in this methodology as we utilise Sequence Alignment to determine distances between trip sequences. Hierarchical clustering is, however, much slower than other clustering algorithms such as k-means clustering. To improve efficiency, this paper calculates the distance matrix based on the different trip sequences observed and then weighted clustering is used to account for the number of times each trip sequence occurs in the data. Trip sequences observed rarely were removed prior to calculating the distance matrix. For a very large number of trip sequences it may be necessary to explore more efficient methods which have been developed for clustering protein sequences for example in Loewenstein et al. (2008).

The third complex process is the Apriori Algorithm used in the Association Rule Mining. Tan et al. (2014, p.346) describe the factors affecting computational complexity. The number of transactions and the average transaction width both have an influence. In our case, these correspond to the total number of Bluetooth devices observed ‘regularly’ and the average number of spatial clusters used per person. The total number of items has an influence, and in our case this is the number of spatially similar clusters. The support and confidence thresholds selected will also influence the complexity as lower thresholds will result in less pruning. Each of these factors is

within the control of the analyst who can adjust the support threshold, the definition of 'regular' travellers or, to a lesser extent, the number of spatially similar clusters to reduce processing times.

5 Case Study

The proposed methodology is applied to real-world data provided by Transport for Greater Manchester (TfGM). Since 2011, TfGM have been installing fixed Bluetooth detectors alongside major arterials and orbitals in and around key urban centres in Greater Manchester such as Manchester, Wigan and Rochdale, for the purpose of monitoring travel times. Antennae with 9dBi gain are used, which Bhaskar and Chung (2013) found provided a range of approximately 100 m. TfGM adjust the strength of detectors on installation to account for the size of the junction. An algorithm is used to truncate and encrypt MAC addresses prior to storing the data. TfGM compared Automatic Number Plate Recognition and Bluetooth detection data for one link over a twelve hour period and calculated hourly penetration rates (of Bluetooth detectors to vehicles) between 16% and 34%.

The case study is limited to eight Bluetooth detectors in and around Wigan town centre (Fig. 6). Data was analysed for a one year period from 1/1/2015 to 31/12/2015. The sensors have all been installed at a similar height although their position relative to

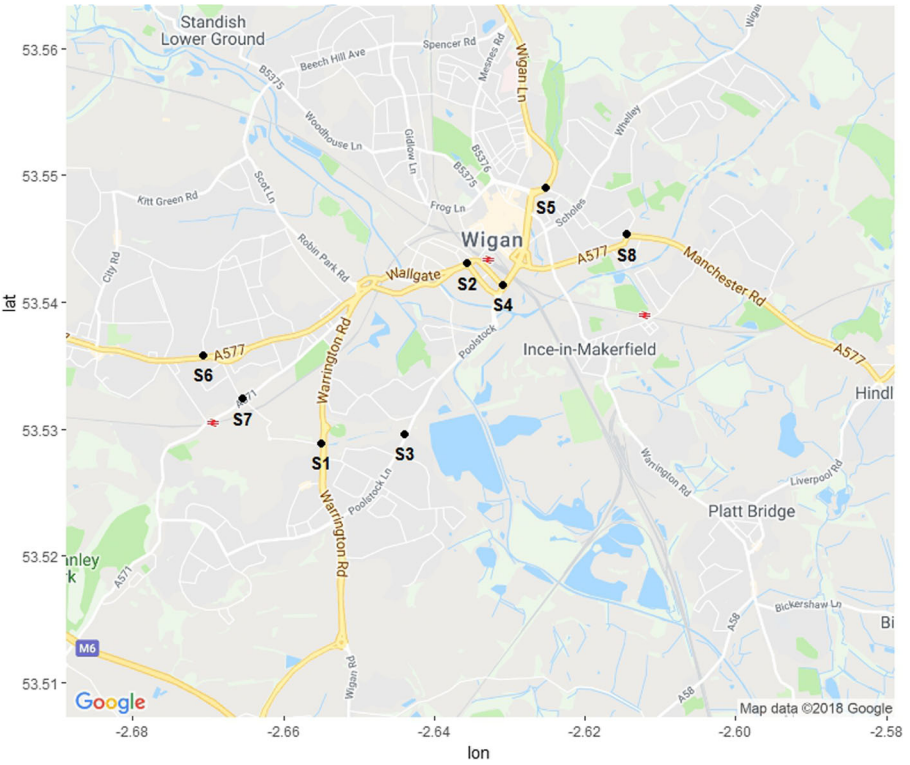

Fig. 6 Map of case study area including Bluetooth detector locations

traffic varies, as will be discussed in Section 5.1. A smaller number of sensors has been used compared with the analysis undertaken in Crawford et al. (2018) so that the results from the association rule mining can be summarised within the space available in this paper and the explanations do not require detailed information about trip attractors and road attributes within different sections of the case study area.

The raw Bluetooth data was processed into trip sequences using the procedures described in Crawford et al. (2018). If the time difference between successive observations of the same traveller is not consistent with driving directly between the two locations then the string is split into separate trips. This decision is made based on the distance between the detectors, the relevant speed limits, and the travel times of surrounding Bluetooth devices. This cleaning process, therefore, seeks to identify trips made within motorised vehicles only. Bluetooth data has been used in other research for pedestrian analysis (Delafontaine et al. 2012; Malinovskiy et al. 2012; Versichele et al. 2012) and cyclist travel times (Mei et al. 2012) but the difficulty arises in using Bluetooth to collect data on multiple modes as it is not usually possible to differentiate between a trip made by car with a stop en-route, and a trip by a slower mode such as cycling.

5.1 Overview and Dimension Reduction

After cleaning and matching data from the eight sites, 2.3 million trips made by 196,557 devices remained. The current research only examines regular travellers, which are defined as devices which recorded 52 or more trips within the case study area during the year. In total, 9564 devices satisfied this criterion and together they recorded 1.4 million trips. As might be expected, the observations of these regular travellers are slightly skewed towards the town centre, with 30% of all of the observations occurring at S4 and 17% at S2.

To apply the methodology proposed in Section 4.1, sensor specific detection probabilities need to be estimated. For this case study, this is done by examining the factors affecting detection rates at other Bluetooth detectors in the Greater Manchester area and then using these relationships to estimate the detection rates for the case study detectors.

Trios of Bluetooth detectors on the same or adjacent links were identified across Greater Manchester where the shortest path between the two outer sensors passes the central sensor. For each trio, the proportion of trips between the outer sensors which were also detected at the central sensor was calculated. Seven trios in Greater Manchester were examined. Most of these locations are not within the case study area, but they involve the same detector type. All available data from 2015 was examined. For each location, the detection rate for each direction of travel was estimated.

Five of the seven trios have detection rates of between 81% and 89% when combining data from both directions. This is consistent with the 80% detection rate found by Araghi et al. (2014). The other trios have much lower detection rates, however. The lowest detection rates were from a site with a substantial difference in the detection rate depending on the direction of travel (66% and 38%). This was the only sensor examined which was not on a straight section of road. The detector is on the outer corner of a fairly sharp bend which has very wide lanes on the inner side of the bend. The detection rates observed at this site are therefore considered to be relatively atypical. The central sensor in this trio is one of the case study detectors

and so the estimated detection probabilities can be used directly within the analyses below. No other case study detectors have similar characteristics in terms of their placement.

Examination of ATC counts and BT detections at the same location suggested that the probability of detection increased as flow increased, although it was not a linear relationship. This could be due to higher traffic volumes resulting in congestion slowing down vehicles, resulting in a longer period of time in the Bluetooth sensor's detection zone and thus a higher probability of being recorded. A quadratic equation between the estimated Bluetooth detection rate and Bluetooth detections per lane was fit to the data from the remaining trios and resulted in an R^2 of 0.96. This relationship was then used to estimate detection probabilities for the case study sites as shown in Table 2.

These detection rates were then used to calculate the sensor-specific indel costs for the Sequence Alignment process. Hierarchical clustering of the trip sequences was then undertaken using the pairwise distances from the Sequence Alignments to populate the distance matrix. The collection of partition quality measures available in the WeightedCluster package (Studer 2013), including the Average Silhouette Width and the Calinski-Harabasz Index, was used to determine that 55 was the optimal number of clusters.

Each cluster contains a set of trip sequences (described by a series of Bluetooth observations). As an example, the most frequently observed trip sequences for two of the clusters are shown in Figs. 7 and 8. When considering sequences observed 50 or more times by any device during the year, the clusters contained 6.5 different sequences on average. The *trips* observed during the year were not evenly distributed amongst the clusters; the largest two clusters contained 28% of trips observed during the year.

To determine the impact of using sensor-specific indel costs, the clustering was also performed using a fixed indel cost for all sensors for comparison purposes. Although the choice of 55 clusters was optimal when using sensor-specific indel costs, this was not the case when using fixed indel costs. In order to have the same number of clusters from the two methods for a more meaningful comparison, the number of clusters selected for the sensitivity analysis was a compromise based on the cluster quality measures from the two sets of analyses. Therefore, for the sensitivity analysis, 60

Table 2 Estimated detection rates for the case study sites

Ref	Total Bluetooth detections	Average daily Bluetooth detections	Total number of lanes	Bluetooth detections per lane (daily)	Estimated detection rate (%)
S1	3,112,373	8527	4	2132	73
S2	2,994,452	8204	4	2051	71
S3	2,605,651	7139	2	3569	88
S4	5,013,170	13,735	6	2289	76
S5	3,339,419	9149	5	1830	66
S6	2,598,288	7119	2	3559	88
S7	1,586,395	4346	2	2173	74
S8	1,598,050	4378	3	1459	57*

*The detection probability for this detector was measured directly

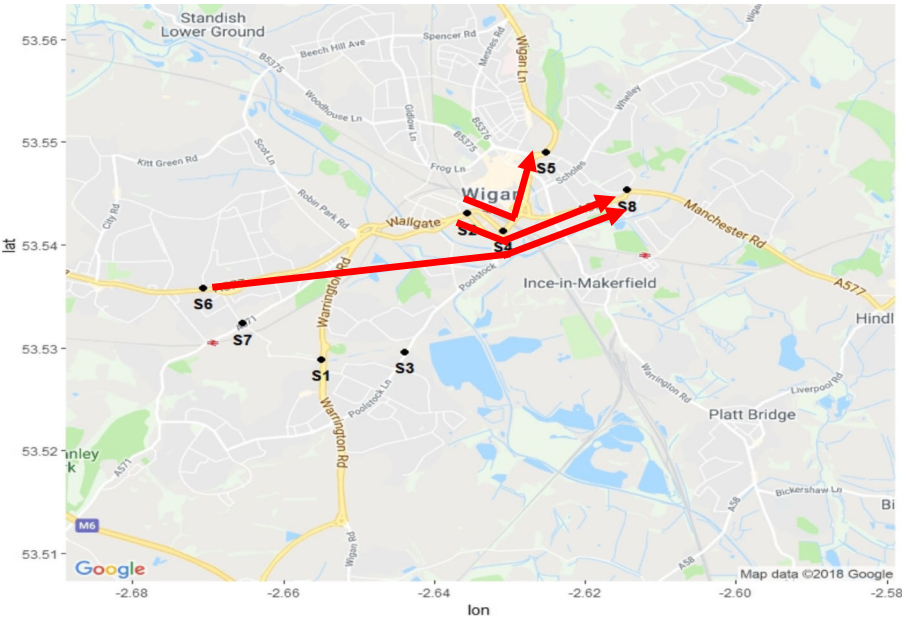

Fig. 7 Sequences in cluster X16 with at least 50 observations during the year

clusters were selected in both cases. Figure 9 compares the cluster membership between the two methods by showing for each of the fixed indel cost clusters, the degree to which those sequences were assigned to the same sensor-specific indel cost cluster. Of the 60 clusters produced using the method proposed in Crawford et al. (2018), 47

Fig. 8 Sequences in cluster X2 with at least 50 observations during the year

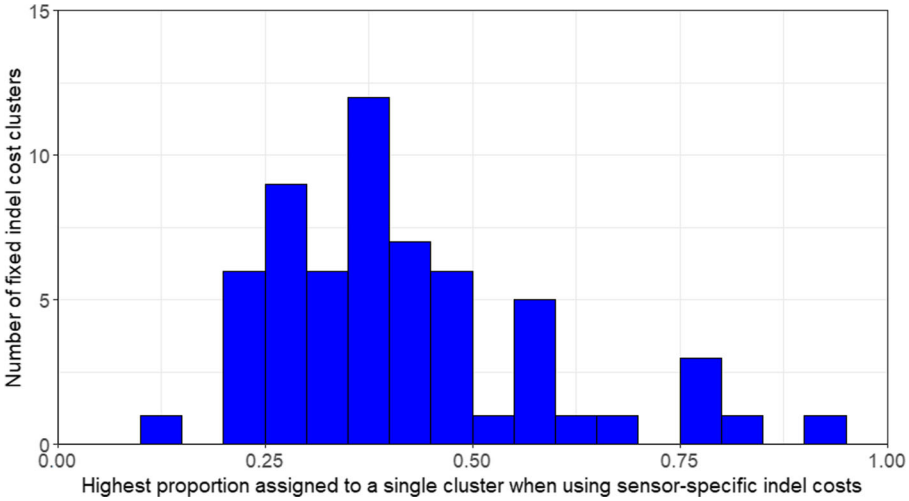

Fig. 9 Highest proportion of sequences in one sensor-specific indel cost cluster for each of the fixed indel cost clusters

clusters had a maximum of 50% of the sequences in that cluster assigned to a single cluster in the revised method, using sensor-specific indel costs as proposed in this paper. This demonstrates that the effort required to estimate sensor-specific detection probabilities and to include them in the Sequence Alignment process is worthwhile as it substantially improved the clusters obtained.

5.2 Spatial Intrapersonal Variability (Traveller Perspective)

On average, the regular travellers made trips within 17 different spatial clusters during the year (Fig. 10). The maximum number of spatial clusters which could have been used was 55 and only 3% of the travellers used 35 or more.

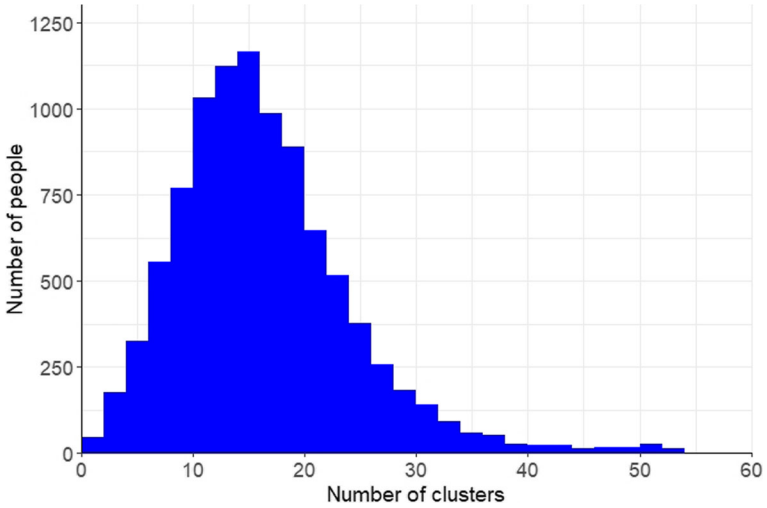

Fig. 10 Histogram showing numbers of spatial clusters per person

The Herfindahl–Hirschman Index (HHI) considers the proportion of trips assigned to each spatial cluster and therefore provides a more comprehensive view of intrapersonal variability. The normalised HHI for each regular traveller is shown in Fig. 11. A value of zero represents an even spread of trips across all spatial clusters and a value of one means that all the traveller's trips belong to the same spatial cluster. For the case study area, this distribution is heavily skewed to the right. The peak has a HHI of approximately 0.08 which could be obtained by using 21 of the spatial clusters during the year and using 5 of those frequently.

One application of these measures of spatial diversity is to examine whether people exhibit more or less spatial diversity at different times, for example according to the day of the week or the time of year. For the case study area, the Herfindahl–Hirschman Index was calculated twice for each person – once for trips on weekdays and once for trips during the weekend. The comparison was not possible for all of the regular travellers as 9% made no trips by car/van in the case study area on weekend days. Also, one traveller made regular trips on weekend days, but none on weekdays during the year. Where the remaining people made at least 10 trips on weekdays and at least 10 trips on weekend days, the Herfindahl–Hirschman Indices were plotted in Fig. 12.

Figure 12 is a heatmap where each square in the grid has a colour based on the number of people with a HHI for their weekday trips within the square's x-axis range and a HHI for their weekend trips within the square's y-axis range. The red line cuts through squares containing people who have weekend and weekday HHIs within the same bin. The people above the red line have higher HHIs for their weekend trips than their weekday trips which means a higher concentration of trips in fewer spatial clusters at the weekend, and vice versa. Of the 6664 people satisfying the minimum sample size criteria, 57% had a larger Herfindahl–Hirschman Index for weekends than weekdays. This suggests

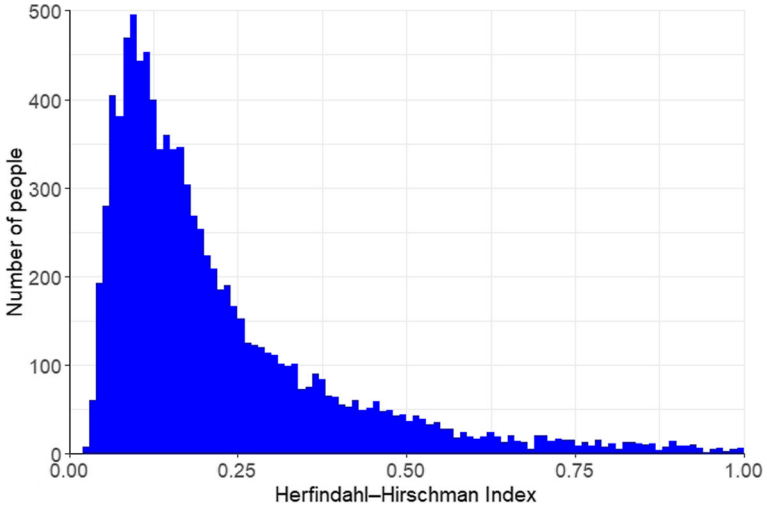

Fig. 11 Histogram of the normalised Herfindahl-Hirschman Index, showing spatial diversity for each person

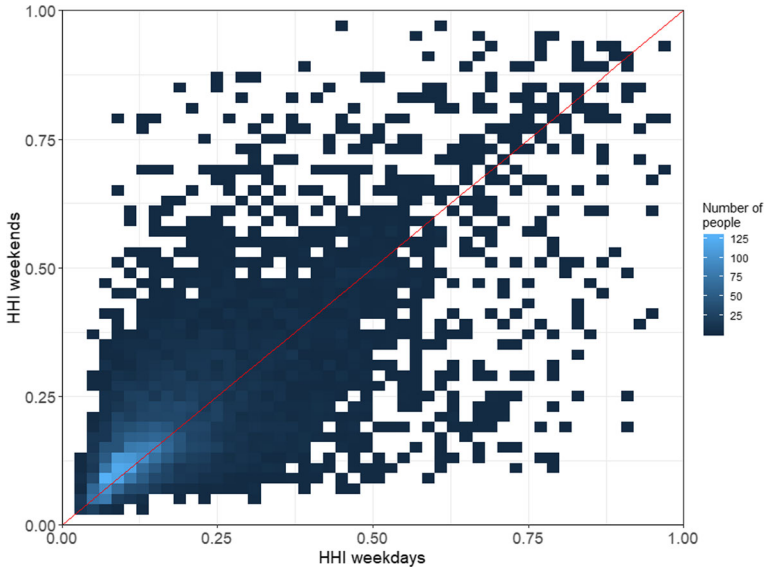

Fig. 12 Comparison of spatial diversity by weekday and weekend days

a slight bias towards people having a more balanced distribution of trips across spatial clusters on weekdays.

Figure 13 compares the HHIs for trips made in the summer and autumn of 2015. No systematic difference in travellers' behaviour between these seasons is observed in terms of spatial diversity.

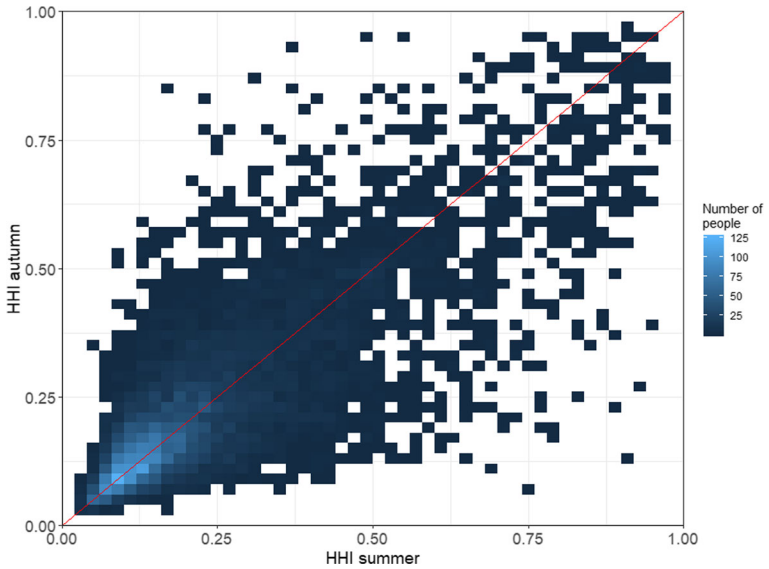

Fig. 13 Comparison of spatial diversity during the summer and autumn of 2015

5.3 Spatial Relationships (Trajectory Perspective)

The association rules with support of at least 0.03 and confidence of at least 0.5 were identified initially. A large number of association rules remained and so to aid interpretation, only the 100 association rules with the highest confidence values were examined in more detail. These rules had confidence values between 0.65 and 0.95. These show high levels of confidence, particularly for the more highly ranked rules, as a confidence of 0.95 means that 95% of the times that rule can be applied, it is correct. The associated values for lift range from 1.67 to 7.56, again indicating a strong relationship particularly for the more highly ranked rules. By including a threshold for support, rules including spatial clusters which are not observed very frequently are excluded. As a result, all of the consequent item sets in the top 100 rules contain the 11 most frequently observed spatially similar clusters.

The most effective way of communicating association rules is through visualisation. Figure 14 includes the 100 association rules with the highest confidence and it was created using the arulesViz package (Hahsler 2019) in R.

In the plot, the clusters of spatially similar trips are denoted by X1 to X55. Each association rule is represented by a circle whose size represents the support for that rule and the colour represents the lift (darker colour represents greater lift). Arrows point from the relevant cluster names to a circle to represent the left hand side of the rule ("if they used this route..."). Arrows point out from the circle to the relevant cluster names representing the right hand side of the rule ("...then they also used this route").

The centre of the plot includes X10, X6 and X4 which are the three clusters most commonly observed on the right hand side of the association rules. These three clusters contain the 2nd, 5th and 3rd most trips, respectively, out of the 55 clusters. All three clusters predominantly contain trip sequences containing just two observations. In the

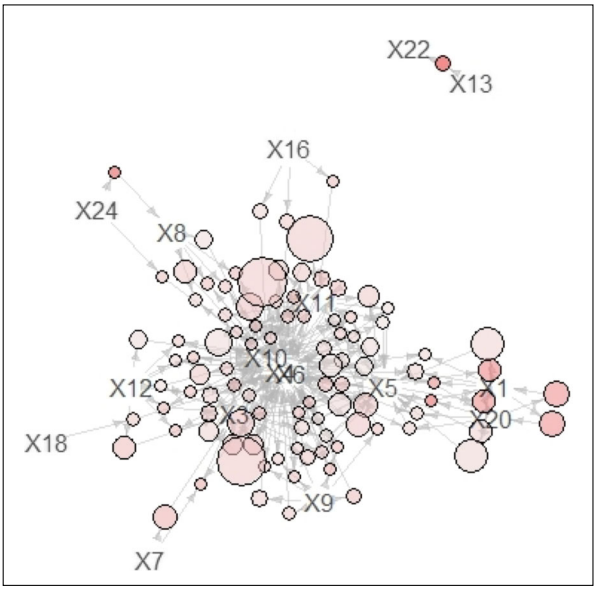

Fig. 14 Graph of the association rules with the highest confidence

case of X10, these are observations in or very close to the town centre. X6 contains short trips *into* the town centre (particularly containing S4) and X4 contains the reverse of those trips. In all, 73 of the association rules have one of these three clusters as the consequent item set and 76 of the 100 rules have at least one of these clusters in the antecedent item set. This is perhaps not surprising as the town centre contains many attractive destinations and short sequences can occur in their own right or as part of longer trips where the Bluetooth device was not detected by other sensors.

In many cases, clusters on the left hand side and the right hand side of rules are pairs in that they include sequences containing the same sensor observations but in the opposite order. This is expected since many outbound trips will be followed by a return trip passing the same sensors but in reverse. What is interesting, however, is that rules with multiple clusters in the antecedent item set typically have higher confidence than related rules with just one cluster in the antecedent item set. For example, clusters X10 and X11 represent short trips in or close to the town centre going westbound and eastbound respectively. The rule $\{X11\} \rightarrow \{X10\}$ has a confidence value of 0.81 and a lift of 2.17. If, however, we consider people who have made trips in cluster X11 and who have also made trips in cluster X9, then the confidence rises to 0.95 and the lift to 2.53. Cluster X9 contains slightly longer trips from the North or East through the town centre.

The other clusters are not equally distributed around the dense central area, however. In the top right hand side of Fig. 14, spatial clusters X22 and X13 sit completely detached from the rest of the graph. The rule $\{X13\} \rightarrow \{X22\}$ is the only rule out of the top 100 association rules which contains either of these clusters. This was the rule with the highest lift value (7.56). It is not surprising that these two clusters are related given that the most common trip sequence in X22 is S7 to S1 and the most common sequence in X13 is S1 to S7 (see Fig. 15). Their disconnect from the other clusters tells us that there is no route which people travelling between S1 and S7 typically use within the town. This does not mean that people using these routes do not use other parts of the network, it just means that there is not a strong enough relationship with any other specific route. Given that these clusters are not related to the commonly used town centre routes, however, this suggests that many of the X22 / X13 travellers do not also visit the town centre. One possible use of such information could be in the placement of signs providing information to travellers in the area, for example about the use of Bluetooth sensors for monitoring travel times and patterns. This analysis suggests that signs in the town centre will be seen at some point by people travelling along many parts of the network, but that an additional sign between S1 and S7 may be required to reach this separate group of people.

On the right hand side of Fig. 14 there is a pair of clusters which are not completely detached from the main graph, but they could be considered to be peripheral. The most common trip sequences in clusters X1 and X20 are shown in Fig. 16. These two sequences pass the same sensors but in the opposite order. Although these two clusters are related to the main body of the graph, the majority of the connection occurs through cluster X5. Cluster X5 predominantly contains trips between S3 and S4 in both directions. These trips are from the town centre going South and they overlap with the southern parts of the sequences in X1 and X20. This provides insights which could be useful for data collection, particularly intercept surveys. Given that clusters X1 and X20 represent trips passing through Wigan from North to South, if the survey aimed to

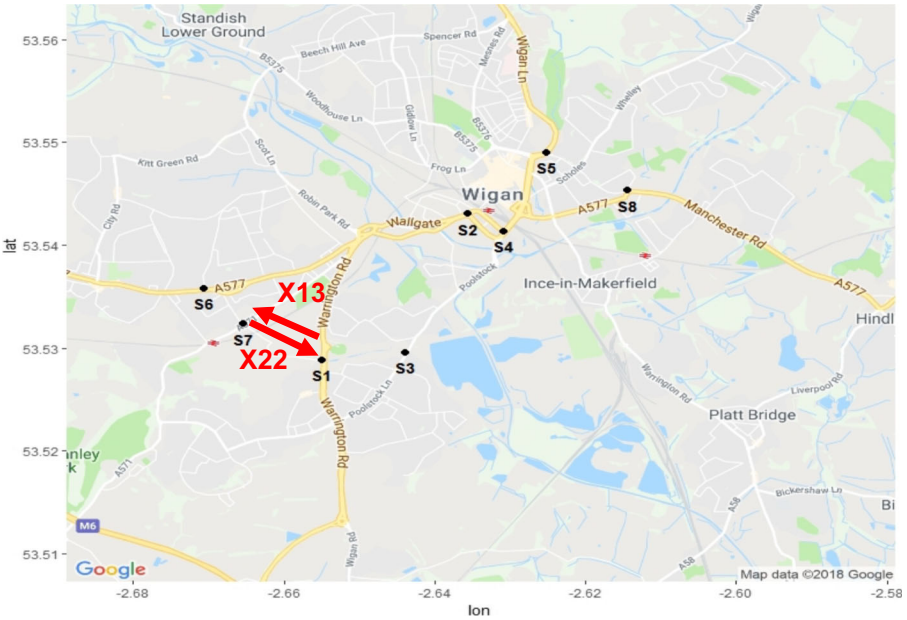

Fig. 15 Most common trip sequences in spatial clusters X13 and X22

capture people passing through in this manner but also people making other types of trips in the area, then somewhere between S3 and S4 (i.e. in cluster X5) might be a suitable location for an intercept survey.

In Fig. 14, the more peripheral rules tend to have relatively low support but high lift. This means that the clusters involved are observed less frequently but that the relationships between the spatial clusters are strong. This need not always be the case with peripheral rules, but in this case they represent more peripheral trip sequences which

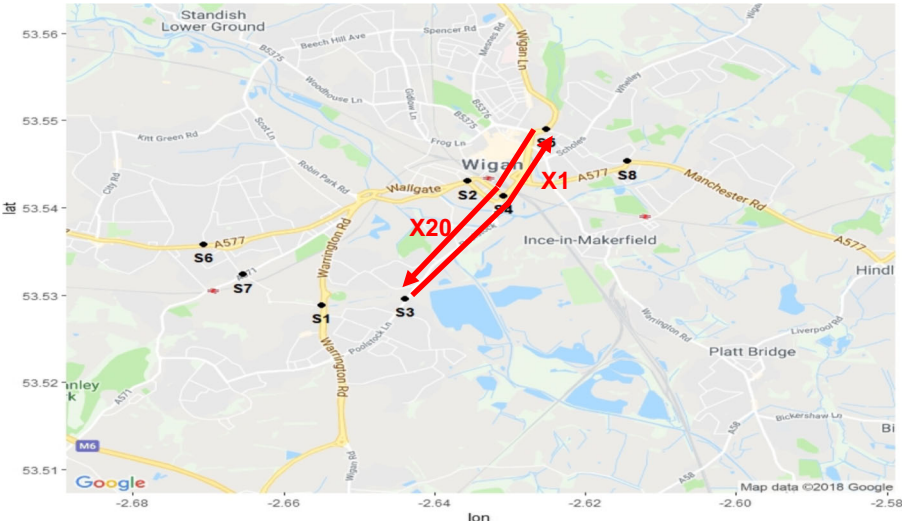

Fig. 16 Most common trip sequences in spatial clusters X1 and X20

may or may not pass through the town centre. Compared with the clusters in the centre of Fig. 14 which are in the town centre, there are likely to be fewer alternative routes available and fewer observed trips.

6 Applications

Section 5 demonstrated how the proposed method could provide insights into the behaviour of regular road users even for a small case study. As well as helping network managers to understand their users better, the methods proposed in this paper could also be used to estimate inputs for other analyses. In this section these further applications will be discussed in terms of modelling, monitoring and evaluation, and service design.

Several existing models account for users with different amounts of information. Most models of this type relate to traveller information systems, such as the models developed by Bifulco et al. (2016) which account for compliance with the information provided in terms of the accuracy of the information and penetration rate of the information systems. These models could be extended to include the network knowledge of different user classes, and the methodology in the current paper could be used to estimate the relevant parameters. Similarly, Li et al. (2017) examine route choice behaviour using user classes with different types of knowledge about network conditions and this could be extended to account for existing network experience. The current research could also inform user classes relating to demand regularity as was proposed in Han et al. (2018).

The type of information discovered about road users using the methods proposed in this paper could also be useful for understanding traveller response to network disruptions. Papangelis et al. (2016, p.63) have highlighted the role which previous experience and knowledge plays in both short and long term impacts of disruptions on traveller behaviour. Current behaviour could also provide insights into travellers' perceived comfort zones (Ngoduy et al. 2013) which could shape their responses. By examining the association rules between disrupted links and alternative routes, road network operators may be able to produce better predictions of how travellers might re-route or choose alternative destinations.

Since habit and inertia also play a role in route switching behaviour (Vacca et al. 2019), for example, it is crucial to be able to quantify the regularity of current behaviour if the network operators wish to affect change.

The outputs from the methods described above may also be useful for calibrating and/or validating multi-day models such as activity based models or day-to-day dynamical models within the framework described in Watling and Cantarella (2015). The methods may also inform new types of models which are likely to evolve based on the availability of 'Big Data' (Milne and Watling 2019). Such models are likely to be more empirically driven and therefore methods such as these which allow us to examine the underlying mechanisms of traveller behaviour (in this case over multiple days) is crucial.

As well as modelling implications, the methodology may also be valuable for monitoring and evaluating network interventions or disruptions. For example, the methodology provides a way of quantifying the extent to which the same travellers will benefit from or be disrupted by two network interventions in the same city. The

information obtained could therefore be used for scheme appraisals and for evaluation studies to explore the equitability of funding across the network. For example, maintenance to remove potholes or gritting in the winter may currently be undertaken on roads which are typically used by the same group of people.

The current research examines the routes through the network which people take and therefore provides a more direct measure of those benefiting or losing out due to investment in the road network. As highlighted by Park and Kwan (2018), examining residential locations only ignores the movement of people through their daily lives. This is equally important when determining whether there is segregation in the use of the road network.

Residential data is also used in the evaluation of transport infrastructure. Dalton et al. (2013), for example, use nearby residential population to measure the beneficiaries of infrastructure such as gyms and parks. Whilst this may be reasonable, they also use this approach for examining the beneficiaries of cycle lanes which is arguably less justifiable. When considering motorised vehicles using the road network, as in the current paper, it is even more important to focus on actual users rather than local residents if such data is available.

The proposed methodology could also be used within evaluation processes to examine how pre- and post-intervention behaviour differs. This would be particularly relevant for policies affecting a zone such as congestion charging and low emission zones which often charge by the day not the trip. In doing so, the results could also be used to verify the results from scheme appraisal modelling (models such as de Palma and Lindsey (2006) and Takama and Preston (2008)).

By providing insights into traveller needs, the methodology presented could also be used to inform real-life system design such as the spatial allocation of on-demand service vehicles or the subsidies provided to Mobility As A Service providers to serve less profitable areas.

7 Conclusions

The case study has demonstrated that even for a small town with just eight detectors, the proposed methodology can be used to measure spatial intrapersonal variability and can provide unexpected insights into the difference in variability between days of the week and seasons. The use of association rules also provides insights into network usage which cannot be obtained from examining traffic counts or an equilibrium model. The outputs from the association rule mining would be useful for stimulating debate about where to target signage, data collection or interventions.

There are limitations stemming from the type of data used. Perhaps most importantly, socio-demographic data is not often available for point-to-point sensor data. This is problematic if the analysis aims to examine issues surrounding equity. Also, when considering issues relating to new services or behaviour change, it is important to understand the demographic characteristics of people making certain trips. For some types of data, particularly ETC and ANPR data, additional data about vehicle owners may be accessible, although perhaps at an additional cost. Where it is not possible to access such data, additional data collection such as intercept surveys or household surveys may

be required to supplement the findings from the analysis based on the methods presented in this paper.

Future work could extend the association rule mining methodology to differentiate between frequent and occasional use of different spatial clusters to provide more nuanced results. Section 6 also discussed many possible applications for the outputs from the proposed methods, but further work is required to integrate outputs from the proposed methodology into modelling frameworks and to demonstrate how the outputs could be communicated to policy makers.

This paper presents an innovative methodology to gain new insights from point-to-point sensor data. Point-to-point sensor data is not new, for example number plates have been recorded in studies for many decades, but the scale and types of data available are growing rapidly. As technological developments rapidly change our mobility patterns, mobility services and the data available on mobility, it is crucial that we continue to develop new methods to gain insights from the available data and our modelling frameworks evolve so that we can understand current behaviour but also shape the future of mobility.

References

Abbott A (1995) Sequence analysis: new methods for old ideas. Annu Rev Sociol 21(1):93–113

Agrawal R, Srikant R (1994) Fast algorithms for mining association rules. In: Proceedings of the 20th VLDC conference (Santiago, Chile) [online]. 1994 pp. 487–499

Agrawal R, Imielinski T, Swami A (1993) Mining association rules between sets of items in large databases. In: Proceedings of the 1993 ACM SIGMOD conference (Washington DC, USA). 1993

Barcelö J, Montero L, Marqués L, Carmona C (2010) Travel time forecasting and dynamic origin-destination estimation for freeways based on Bluetooth traffic monitoring. Transp Res Rec 2175(1):19–27

Bayarma A, Kitamura R, Susilo Y (2007, 2021) Recurrence of daily travel patterns: stochastic process approach to multiday travel behavior. Trans Res Rec: J Trans Res Board:55–63

Bhaskar A, Chung E (2013) Fundamental understanding on the use of Bluetooth scanner as a complementary transport data. Trans Res Part C Emerg Technol 37:42–72

Bifulco GN, Cantarella GE, Simonelli F, Velonà P (2016) Advanced traveller information systems under recurrent traffic conditions: network equilibrium and stability. Transp Res B Methodol 92:73–87

Buliung RN, Roorda MJ, Remmel TK (2008) Exploring spatial variety in patterns of activity-travel behaviour: initial results from the Toronto travel-activity panel survey (TTAPS). Transportation 35(6):697

Calabrese F, Colonna M, Lovisolo P, Parata D, Ratti C (2011) Real-time urban monitoring using cell phones: a case study in Rome. IEEE Trans Intell Transp Syst 12(1):141–151

Cantarella GE, Cascetta E (1995) Dynamic processes and equilibrium in transportation networks: towards a unifying theory. Transp Sci 29(4):305–329

Carpenter C, Fowler M, Adler TJ (2012) Generating route-specific origin–destination tables using Bluetooth technology. Transp Res Rec 2308(1):96–102

Chen H, Yang C, Xu X (2017) Clustering vehicle temporal and spatial travel behavior using license plate recognition data. J Adv Transp:2017
Crawford F, Watling DP, Connors RD (2018) Identifying road user classes based on repeated trip behaviour using Bluetooth data. Transp Res A Policy Pract 113:55–74
Dalton AM, Jones A, Ogilvie D, Petticrew M, White M, Cummins S (2013) Using spatial equity analysis in the process evaluation of environmental interventions to tackle obesity: the healthy towns programme in England. Int J Equity Health 12(1):43
de Palma A, Lindsey R (2006) Modelling and evaluation of road pricing in Paris. Transp Policy 13(2):115–126
Delafontaine, M., Versichele, M., Neutens, T. and de Weghe, N. Van (2012) Analysing spatiotemporal sequences in Bluetooth tracking data. *Applied Geography*. 34 pp. 659–668
Dharmowijoyo DBE, Susilo YO, Karlström A (2017) Analysing the complexity of day-to-day individual activity-travel patterns using a multidimensional sequence alignment model: A case study in the Bandung Metropolitan Area, Indonesia. J Transp Geogr 64:1–12
Diana M, Pirra M (2016) A comparative assessment of synthetic indices to measure multimodality behaviours. Transportmetrica A: Trans Sci 12(9):771–793
Dijst M (1999) Two-earner families and their action spaces: A case study of two dutch communities. GeoJournal 48(3):195
Elango V, Guensler R, Ogle J (2007) Day-to-day travel variability in the commute Atlanta, Georgia, study. Trans Res Re: J Transp Res Board 2014:39–49
Felstead A (2012) Rapid change or slow evolution? Changing places of work and their consequences in the UK. J Transp Geogr 21:31–38
Gabadinho A, Ritschard G, Müller NS, Studer M (2011a) Analyzing and visualizing state sequences in R with TraMineR. J Stat Softw 40(4)
Gabadinho, A., Ritschard, G., Studer, M., Müller, N.S. (2011b) Mining sequence data in R with the TraMineR package: A user's guide (v1.8) [online]
Goulet-Langlois G, Koutsopoulos HN, Zhao J (2016) Inferring patterns in the multi-week activity sequences of public transport users. Transp Res Part C: Emerg Technol 64:1–16
Hahsler M (2019) arulesViz: visualizing association rules and frequent itemsets [online]
Hainen AM, Wasson JS, Hubbard SML, Remias SM, Farnsworth GD, Bullock DM (2011) Estimating route choice and travel time reliability with field observations of Bluetooth probe vehicles. Transp Res Rec 2256(1):43–50
Han L, Sun H, Wang DZW, Zhu C (2018) A stochastic process traffic assignment model considering stochastic traffic demand. Transportmetrica B: Transport Dynamics 6(3):169–189
Haseman RJ, Wasson JS, Bullock DM (2010) Real-time measurement of travel time delay in work zones and evaluation metrics using Bluetooth probe tracking. Transp Res Rec 2169(1):40–53
Heinen E, Chatterjee K (2015) The same mode again? An exploration of mode choice variability in Great Britain using the National Travel Survey. Transp Res A Policy Pract 78:266–282
Hirsch JA, Winters M, Ashe MC, Clarke PJ, McKay HA (2016) Destinations that older adults experience within their GPS activity spaces: relation to objectively measured physical activity. Environ Behav 48(1): 55–77
Houston D, Luong TT, Boarnet MG (2014) Tracking daily travel; assessing discrepancies between GPS-derived and self-reported travel patterns. Transportation Research Part C: Emerging Technologies 48:97–108
Huff JO, Hanson S (1986) Repetition and variability in urban travel. Geogr Anal 18(2):97–114
Järv O, Ahas R, Witlox F (2014) Understanding monthly variability in human activity spaces: a twelve-month study using mobile phone call detail records. Transp Res C: Emerg Technol 38:122–135
Joh C-H, Arentze TA, Timmermans HJP (2001) A position-sensitive sequence-alignment method illustrated for space–time activity-diary data. Environment and Planning A: Economy and Space 33(2):313–338
Joh C-H, Arentze T, Hofman F, Timmermans H (2002) Activity pattern similarity: a multidimensional sequence alignment method. Transp Res B Methodol 36(5):385–403
Jones P, Clarke M (1988) The significance and measurement of variability in travel behaviour. Transportation. 15(1):65–87
Kieu LM, Bhaskar A, Chung E (2015) Passenger segmentation using smart card data. IEEE Trans Intell Transp Syst 16(3):1537–1548
Kim J, Mahmassani HS (2015) Spatial and temporal characterization of travel patterns in a traffic network using vehicle trajectories. Transportation Research Procedia 9:164–184
Kim J, Kurauchi F, Uno N, Hagihara T, Daito T (2014) Using electronic toll collection data to understand traffic demand. J Intell Transp Syst 18(2):190–203

Kim J, Corcoran J, Papamanolis M (2017) Route choice stickiness of public transport passengers: measuring habitual bus ridership behaviour using smart card data. Transportation Research Part C: Emerging Technologies 83:146–164

Kurauchi F, Schmöcker J-D, Shimamoto H, Hassan SM (2014) Variability of commuters' bus line choice: an analysis of oyster card data. Public Transport 6(1):21–34

Li M, Rouphail NM, Mahmoudi M, Liu J, Zhou X (2017) Multi-scenario optimization approach for assessing the impacts of advanced traffic information under realistic stochastic capacity distributions. Transportation Research Part C: Emerging Technologies 77:113–133

Liu X, Gong L, Gong Y, Liu Y (2015) Revealing travel patterns and city structure with taxi trip data. J Transp Geogr 43:78–90

Loewenstein Y, Portugaly E, Fromer M, Linial M (2008) Efficient algorithms for accurate hierarchical clustering of huge datasets: tackling the entire protein space. Bioinformatics. 24(13):i41–i49

Malinovskiy Y, Saunier N, Wang Y (2012) Analysis of pedestrian travel with static Bluetooth sensors. Transp Res Rec 2299(1):137–149

Masso A, Silm S, Ahas R (2019) Generational differences in spatial mobility: a study with mobile phone data. Popul Space Place 25(2):e2210

McLeod FN, Cherrett TJ, Box S, Waterson BJ, Pritchard JA (2017) Using automatic number plate recognition data to investigate the regularity of vehicle arrivals. Eur J Transp Infrastruct Res 17(1):86–102

Mei Z, Wang D, Chen J (2012) Investigation with Bluetooth sensors of bicycle travel time estimation on a short corridor. International Journal of Distributed Sensor Networks 8(1):303521

Millward H, Hafezi MH, Daisy NS (2019) Activity travel of population segments grouped by daily time-use: GPS tracking in Halifax, Canada. Travel Behav Soc 16:161–170

Milne D, Watling D (2019) Big data and understanding change in the context of planning transport systems. J Transp Geogr 76:235–244

Moghaddam SS, Hellinga B (2013) Quantifying measurement error in arterial travel times measured by Bluetooth detectors. Transp Res Rec 2395(1):111–122

Moya-Gómez B, Salas-Olmedo MH, García-Palomares JC, Gutiérrez J (2018) Dynamic accessibility using big data: the role of the changing conditions of network congestion and destination attractiveness. Netw Spat Econ 18(2):273–290

Muthyalagari GR, Parashar A, Pendyala RM (2001) Measuring day-to-day variability in travel characteristics using GPS data. In: Proceedings from the 80th Annual Meeting of the Transportation Research Board. 2001

Ngoduy D, Watling D, Timms P, Tight M (2013) Dynamic Bayesian belief network to model the development of walking and cycling schemes. Int J Sustain Transp 7(5):366–388

Pande A, Abdel-Aty M (2009) Market basket analysis of crash data from large jurisdictions and its potential as a decision support tool. Saf Sci 47(1):145–154

Papangelis K, Velaga NR, Ashmore F, Sripada S, Nelson JD, Beecroft M (2016) Exploring the rural passenger experience, information needs and decision making during public transport disruption. Res Transp Bus Manag 18:57–69

Park YM, Kwan M-P (2018) Beyond residential segregation: a spatiotemporal approach to examining multi-contextual segregation. Comput Environ Urban Syst 71:98–108

R Core Team (2019) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.r-project.org/

Ramaekers K, Reumers S, Wets G, Cools M (2013) Modelling route choice decisions of Car Travellers using combined GPS and diary data. Netw Spat Econ 13(3):351–372

Safi H, Assemi B, Mesbah M, Ferreira L, Hickman M (2015) Design and implementation of a smartphone-based travel survey. Transportation Research Record: Journal of the Transportation Research Board 2526: 99–107

Schlich R, Schönfelder S, Hanson S, Axhausen KAYW (2004) Structures of leisure travel: temporal and spatial variability. Transp Rev 24(2):219–237

Schönfelder S, Axhausen KW (2003) Activity spaces: measures of social exclusion? Transp Policy 10(4):273–286

Schuessler N, Axhausen KW (2009) Map-matching of GPS traces on high-resolution navigation networks using the multiple hypothesis technique (MHT). October. (October), pp. 1–22

Shao H, Lam WHK, Tam ML (2006) A reliability-based stochastic traffic assignment model for network with multiple user classes under uncertainty in demand. Netw Spat Econ 6(3):173–204

Shen Y, Kwan M-P, Chai Y (2013) Investigating commuting flexibility with GPS data and 3D geovisualization: a case study of Beijing, China. J Transp Geogr 32:1–11

Shen B, Zheng W, Carley KM (2018) Urban activity mining framework for ride sharing systems based on vehicular social networks. Netw Spat Econ 18(3):705–734
Shoval N, Isaacson M (2007) Sequence alignment as a method for human activity analysis in space and time. Ann Assoc Am Geogr 97(2):282–297
Stock K (2018) Mining location from social media: a systematic review. Comput Environ Urban Syst 71:209–240
Stopher PR, Zhang Y (2011) Repetitiveness of daily travel. Transp Res Rec 2230(1):75–84
Studer M (2013) WeightedCluster library manual: A practical guide to creating typologies of trajectories in the social sciences with R [online]
Susilo YO, Axhausen KW (2014) Repetitions in individual daily activity–travel–location patterns: a study using the Herfindahl–Hirschman index. Transportation. 41(5):995–1011
Susilo YO, Kitamura R (2005) Analysis of day-to-day variability in an Individual's action space: exploration of 6-week Mobidrive travel diary data. Transp Res Rec 1902(1):124–133
Takama T, Preston J (2008) Forecasting the effects of road user charge by stochastic agent-based modelling. Transp Res A Policy Pract 42(4):738–749
Tam ML, Lam WHK (2008) Using automatic vehicle idenification data for travel time estimation in Hong Kong. Transportmetrica. 4(3):179–194
Tan P-N, Steinbach M, Kumar V (2014) Introduction to data mining. Pearson, Harlow, Essex
Toch E, Lerner B, Ben-Zion E, Ben-Gal I (2019) Analyzing large-scale human mobility data: a survey of machine learning methods and applications. Knowl Inf Syst 58(3):501–523
Traunmueller MW, Johnson N, Malik A, Kontokosta CE (2018) Digital footprints: using WiFi probe and locational data to analyze human mobility trajectories in cities. Comput Environ Urban Syst 72:4–12
Tu W, Cao R, Yue Y, Zhou B, Li Q, Li Q (2018) Spatial variations in urban public ridership derived from GPS trajectories and smart card data. J Transp Geogr 69:45–57
Vacca A, Prato CG, Meloni I (2019) Should I stay or should I go? Investigating route switching behavior from revealed preferences data. Transportation. 46(1):75–93
van Heeswijck T, Paquet C, Kestens Y, Thierry B, Morency C, Daniel M (2015) Differences in associations between active transportation and built environmental exposures when expressed using different components of individual activity spaces. Health Place 33:195–202
Versichele M, Neutens T, Delafontaine M, Van de Weghe N (2012) The use of Bluetooth for analysing spatiotemporal dynamics of human movement at mass events: a case study of the Ghent festivities. Appl Geogr 32(2):208–220
Wang JYT (2015) 'Resilience thinking' in transport planning. Civ Eng Environ Syst 32(1–2):180–191
Wang Z, He SY, Leung Y (2018) Applying mobile phone data to travel behaviour research: a literature review. Travel Behav Soc 11:141–155
Watling DP, Cantarella GE (2015) Model representation & decision-making in an ever-changing world: the role of stochastic process models of transportation systems. Netw Spat Econ 15(3):843–882
Watling D, Milne D, Clark S (2012) Network impacts of a road capacity reduction: empirical analysis and model predictions. Transp Res A Policy Pract 46(1):167–189
Wilson WC (1998) Activity pattern analysis by means of sequence-alignment methods. Environment and Planning A: Economy and Space 30(6):1017–1038
Yoshimura Y, Sobolevsky S, Hobin JNB, Ratti C, Blat J (2018) Urban association rules: uncovering linked trips for shopping behavior. Environ Plan B Urban Anal City Sci 45(2):367–385
Zhan X, Ukkusuri SV, Zhu F (2014) Inferring urban land use using large-scale social media check-in data. Netw Spat Econ 14(3–4):647–667