Multimodal Transportation 2 (2023) 100049

Contents lists available at ScienceDirect

Multimodal Transportation

journal homepage: www.elsevier.com/locate/multra

Full Length Article

Investigating pedestrian behaviour in urban environments: A Wi-Fi tracking and machine learning approach

Avgousta Stanitsa*, Stephen H Hallett, Simon Jude

Cranfield Environment Centre, School of Water, Energy and Environment, Cranfield University, Cranfield, UK, Bedfordshire, MK43 0AL

ARTICLE INFO

Keywords:
Pedestrian movement
Machine-learning
Urban environment
Wi-Fi tracking
Human behaviour

ABSTRACT

Urban geometry plays a critical role in determining paths for pedestrian flow in urban areas. To improve the urban planning processes and to enhance quality of life for end-users in urban spaces, a better understanding of the factors influencing pedestrian movement is required by decision-makers within the urban design and planning industry. The aim of this study is to present a novel means to assess pedestrian routing in urban environments. As a unique contribution to knowledge and practice, this study: (a) enhances the body of knowledge by developing a conceptual model to assess and classify pedestrian movement behaviours, utilising machine learning algorithms and location data in conjunction with spatial attributes, and (b) extends previous research by revealing spatial visibility as a driver for pedestrian movement in urban environments. The importance of the findings lies in the perspective of revealing novel insights concerning individual preferences and behaviours of end-users and the utilisation of urban spaces. The approaches developed can be utilised for observations in large-scale contexts, as an addition to traditional methods. Application of the model in a high pedestrian traffic-dense retail urban area in London reveals clear and consistent relationships amongst spatial visibility, individuals' motivation, and knowledge of the area. Key behaviours established in the study area are grouped into two activity categories: (i) Utilitarian walking (with motivation - expert and novice striders) and (ii) Leisure walking (no motivation - expert and novice strollers). The approach offers an insightful and automated means to understand pedestrian flow in urban contexts and informs wider wayfinding, walkability, and transportation knowledge.

Introduction

Open urban spaces represent an important asset within cities, providing opportunities for users to engage with their communities and enhance their quality of life (Mouratidis, 2021). Nevertheless, urban growth and development have pressured public urban spaces and, subsequently, their design. The planning and design of a city are influenced by several factors, with mobility being one of the most influential (Mendiola and González, 2021). Movements within street networks and the act of walking aid planners implement road

Abbreviations: ML, Machine Learning; BDAs, Big Data Approaches; TFL, Transport for London; AI, Artificial Intelligence; UWB, Ultra-wideband transceivers; CCTV, Closed-circuit television; MAC, Media Access Control; VGA, Visibility Graph Analysis; GIS, Geospatial Information Systems; HPC, High-Performance Computer; GDPR, General Data Protection Regulation; NaN, Not a Number; EM, Elbow Method; SA, Silhouette Analysis; WCSS, Within-Cluster Sum-of-Squares; CH, Calinski-Harabasz coefficient.

* Corresponding author.

E-mail addresses: avgousta.stanitsa@cranfield.ac.uk, augusta.stanitsa@outlook.com (A. Stanitsa), s.hallett@cranfield.ac.uk (S.H. Hallett), s.jude@cranfield.ac.uk (S. Jude).

https://doi.org/10.1016/j.multra.2022.100049
Received 27 March 2022; Received in revised form 26 May 2022; Accepted 28 June 2022

development, public transport, and placement of amenities, and create designs promoting physical activity, healthier communities, and well-being (Järv et al., 2012; O'Sullivan et al., 2000; Brown et al., 2014). Although diversity in walking activities is generally recognised in wayfinding and walkability studies (Lynch, 1960; Cornell et al., 2003; Bitgood, 2010), there is not yet a systematic way to best categorise pedestrian behaviours, leaving a gap in knowledge. While there is a large body of research on the effect of the built environment on walking patterns from various perspectives (Gehl, 2011; Zacharias, 2001; Clifton et al., 2007; Mehta, 2009), less is known about how and to what extent it influences pedestrian behaviour.

Information is sought by urban planners and designers to aid an understanding of the impact of the built environment on pedestrian behaviour. However, there remains a lack of objective data in the study of pedestrian movement patterns, and consequently, there is a limited understanding of the role and value of implementing novel technologies in design. Increased urban data availability has renewed the interest in urban mobility for better understanding pedestrian activity and the effects of diverse physical factors on behaviour. Nevertheless, to date, studies exploring human experience on urban spaces are mainly based on traditional data sources and analysis methods, such as observational data and statistical techniques (Hillier et al., 1993; Krizek et al., 2009). Although traditional qualitative methodologies have several advantages, including data richness and validity, they present several limitations in terms of controllability, data quality, representativeness, and associated costs (Feng et al., 2021). Such limitations include the difficulty of recording crowd movements in public spaces via observational data and experiments containing bias (Feng et al., 2021).

Existing studies have illustrated the necessity of both contemporary data collection and analysis methods, such as objective walking patterns from large-scale monitoring and machine learning analysis techniques, to enable the study of new types of pedestrian movement (Feng et al., 2021; Lee, 2020). These methods hold the prospect for the collection of new types of pedestrian movement data due to their several advantages, such as increased experimental control and lower implementation costs. They subsequently help to overcome some of the limitations found on traditional approaches (e.g., surveys or observational data collection). Although such approaches highlight the potential of data-driven methodologies in supporting more informed design decisions, it is yet unclear in what way and how novel sources of information and approaches could be realised to provide insights on pedestrian movement behaviours in urban spaces, leaving an additional gap.

These studies make substantial progress in translating great amounts of data and spatial information from cities into specific pedestrian knowledge (e.g., (Zhang et al., 2020; Karbovskii et al., 2019)). However, most studies fail to investigate behaviour in large contexts by focusing on small-scale data samples, bivariate analysis; not multifaceted, manually annotated training data which can be proven costly, or individual detection; which is not systematic, for supervised learning (Koh et al., 2020). As a result, unsupervised pedestrian movement data evidence, such as relationships between pedestrians and other urban objects, is lacking. Employment of large-scale monitoring via smartphones and sensor networks, machine learning and evolutionary computation are becoming prominent in the domain of pedestrian mobility to complement traditional methods for extracting and improving the semantics of human movement behaviour (Wirz et al., 2013).

The aim of this study is to present a novel means to assess pedestrian routing in urban environments. A conceptual model is presented to classify behaviours and spatial configuration interactions, utilising machine learning (ML) algorithms and location data derived from Wi-Fi tracking techniques. The proposed model illustrates numerous differences and advantages compared to other methods in the existing literature. The conceptual model developed utilises large-scale data samples, while existing literature in the study of urban space recognition and the role of sensorial experience on movement patterns is limited to small-scale data samples. To overcome the limitation of lacking labelled data, un-supervised ML clustering is employed utilising data collected from Wi-Fi tracking techniques combined with urban design attributes, and more specifically, spatial visibility as extracted from space syntax methodologies. Thus, the results reveal novel insights concerning individual preferences and behaviours of end-users and the utilisation of urban spaces, which in existing literature are achieved by the collection of qualitative data, such as on-site observations. Finally, the proposed model developed provides a systematic way to assess pedestrian behaviour utilising novel sources of information and approaches in large-scale contexts, covering the existing literature gaps.

As a unique contribution to knowledge and practice, this study: (a) enhances the body of knowledge by developing a conceptual model to assess and classify pedestrian movement behaviours, utilising machine learning algorithms and location data in conjunction with spatial attributes, and (b) extends previous research by revealing the spatial visibility aspect as a driver for pedestrian movement in urban environments. The importance of the findings lies in the perspective of revealing novel insights concerning individual preferences and behaviours of end-users and the utilisation of urban spaces.

Pedestrian behaviour and spatial production

Several studies have been conducted in the field of built environment-pedestrian behaviour relationships from various disciplines over the last three decades (Lynch, 1960; Dridi, 2015; Gehl, 2011; Gibson, 1988). Due to the complexity of pedestrian movement, the approaches suggested to explain it in urban space and the focus of research varies among the fields. For example, in the fields of health and urban design, the emphasis is placed on the qualities and attributes of urban design, treated with reference to the immediate condition of individual streets. Such studies have documented relations among street-level design, pedestrian activity, and environmental correlates of walking (Loukaitou-Sideris, 2020). Research in transportation and planning though, turned its attention to urban form aspects of walkability (i.e., proximity and distance) and connectivity to reveal their relations with pedestrian movement behaviour (Frank, 2000). A thorough literature review on pedestrian behaviour in these fields reveals several important observations. The following outcomes, methods and technologies used in the research to achieve their objectives are presented below:

Findings from observations

In 1970, sociologist William H. Whyte founded The Street Life Project; a small research group that observed many plazas and small parks in New York City to determine the factors that explain how some city spaces respond well to people's needs while others do not, and then documented what might be the basic elements of a successful small urban space (Whyte, 1980). Whyte observed how people seek more than mere physiological comfort, and how pedestrians will consequently undergo a certain degree of physical discomfort to satisfy psychological needs. Distractions and pleasures divert the pedestrian's attention away from intentions to reduce distance and effort, even if reducing effort is related with carrying heavy goods (Al-Widyan et al., 2017; Gärling and Gärling, 1988).

Research to date highlighted the significance of urban space recognition and the role of sensorial experience on movement patterns (Whyte, 1980; Choi, 2012), however, the understanding on the way they are related to spatial behaviour remains limited, with the greater sensory field receiving little attention (Zacharias, 2001). Examples investigating these aspects are recent studies around streetscape features relating to comfort and pleasurability (Capitanio, 2019), emotions against the background of environmental information (Resch et al., 2020) or familiar and unfamiliar spaces (Phillips et al., 2013). These studies highlight that pedestrian behaviour changes differ based on diverse parameters relating to the physiological characteristics of the individuals. Emotional responses are not only part of the individual or collective subjective experiences but constitute a motivational factor for behaviour and choice.

Research investigating the significance of preference and emotional qualities to pedestrian walking patterns, mainly utilises manual gathering of observational and qualitative data. Whyte's work (Whyte, 1980) was seminal, being one of the first attempts to quantify human activity in open spaces using data-driven approaches, via interviews, observations, and the use of a time lapse camera with a digital clock overlooking the plazas to record daily patterns. Employment of traditional qualitative approaches for data collection and analysis presents several advantages, such as data richness and validity (Feng et al., 2021). Field observations and surveys can be conducted over long periods of time, collecting specific characteristics of pedestrians, such as sex, direction, personal items, clothing information, psychological insights (preferences, motivations etc.) and others, resulting in rich and detailed information considering fundamental concepts of influence of human behaviour. In addition, pedestrians do not have the knowledge of being tracked, hence, their response to urban settings is in a more natural fashion. The analysis techniques of traditional data sources rely mainly on statistical models, in which designers and urban planners have been traditionally trained to undertake such tasks (statistics, survey research and estimations) (French et al., 2015). Therefore, such approaches can ensure their practical application in academia, without minimising their research potential.

However, several disadvantages around controllability, data quality, representativeness and associated costs of these methods exist. These approaches remain to-date time consuming and labour-intensive, often limiting the scope of research (Feng et al., 2021). In addition, the accuracy of behavioural data relies on the setup, or the techniques used, such as granularity of the data, mechanisms of recording the data and distribution of their densities, or respondents' internal characteristics, such as past experiences or personal views, resulting in often unreliable datasets, not suitable for detailed analyses. Existing studies have illustrated the necessity of contemporary data collection and analysis methods, whilst highlighting a lack of novel techniques employed in the fields of sensory and urban design, demonstrating the current limitations and disadvantages concerning the types of pedestrian behaviour that can be studied with traditional approaches (Feng et al., 2021). For example, recording concurrent crowd movements in public spaces via observational data is difficult, introducing biased information collected via experimental setups.

Pedestrian activities categorisation

Several researchers attempted to categorise pedestrian activities and their influencing factors. Gehl (Gehl, 2011) simplified outdoor activities in urban spaces to three categories: necessary, optional, and social. Gehl argued that when the quality of the urban environment is good, optional activities increase in frequency. As those activities rise, so the number of social activities also increase (Gehl and Gemzoe, 1996). Other researchers have divided the types of activities to two categories, driven by the pedestrian's motivations, as an example the study of Ki and Lee (Ki and Lee, 2021) where they divided activities into utilitarian and leisure, whilst others have followed similar categorisation, further explained in Table 1. The contribution to knowledge of Table 1 to existing body of literature is to reveal the lack of a systematic knowledge about how to best categorise pedestrian behaviours.

Wayfinding research findings

The extensive literature on wayfinding typically supports the notion that the complexity of spatial design is connected to success in reaching a destination (Dridi, 2015; Weisman, 1981; Gibson, 1988). Literature on wayfinding identifies how within complex built environments there are two key types of journeys: (i) goal-oriented and (ii) non-goal-directed or exploratory (Gibson, 1988). The first refers to pedestrian's motivation in moving towards specific points within a space, e.g., residential buildings or transit terminals. The second is stimulated by visually attractive objects encountered along the path to the goal, e.g., window displays or street performances. According to Transport for London's (TFL) research, there are four different and distinct types of journeys, each with specific travel characteristics, thus: Novice strider, Expert strider, Novice stroller and Expert stroller (Davies, 2007). Within this categorisation,

Table 1
Summary of walking activities categorisation in research.

Walking activities categories	Source	Title	Date	Method	Data source
• Utilitarian walking • Leisure walking	Ki, D. and Lee, S. (Ki and Lee, 2021)	Analyzing the effects of Green View Index of neighborhood streets on walking time using Google Street View and deep learning	2021	Green View Index (GVI), Semantic segmentation, deep neural network model fully convolutional network	Google Street View (GSV) images
• Walking for transport • Walking for recreation	Zhang X., Melbourne S., Sarkar C., Chiaradia A., Webster C. (Zhang et al., 2020)	Effects of green space on walking: Does size, shape, and density matter?	2020	Regression model and statistical analysis	Green spaces from UKMap, Pedestrian data from London Travel Demand Survey 2009/2010
• Essential trips for commuting • Optional trips for recreational activities	Lee, J.M. (Lee, 2020)	Exploring Walking Behavior in the Streets of New York City Using Hourly Pedestrian Count Data	2020	Placemeter utilising computer vision algorithms. Produced video feeds from streets and creates automated reports tracking the number of pedestrians.	Pedestrian count and relative speed data captured by video, weather data from Central Park weather station (KNYC) and other private weather stations in New York City, and sunlight and wind simulation results from massing models of the corresponding locations.
• Stationery activities • Passer-by activities Cultural and social activities	Istrate et al. (Istrate et al., 2020)	How Attractive for Walking Are the Main Streets of a Shrinking City	2020	Case study approach	Observational data
• Optional Necessary • Social (Resultant activities)	Jan Gehl & Birgitte Svarre (Gehl and Svarre, 2013)	How to Study Public Life	2013	Systematic study	Qualitative data, including Observations, interviews, mapping existing infrastructure
• Walking for exercise • Walking for sports/ exercise Walking for transport	Tudor-Locke, Bittman, Merom, D. (Tudor-Locke et al., 2005)	Patterns of walking for transport and exercise: a novel application of time use data	2005	Nesting 3-digit code classification and statistical analysis	Time use diaries
• Goal- oriented • Non-goal-directed or exploratory	Gibson, E.J. (Gibson, 1988)	Exploratory behavior in the development of perceiving, acting, and the acquiring of knowledge	1988	Experiments	Observational data

knowledge of the area is incorporated for all types of activities. These studies highlight that built environment design can either help or hinder individual wayfinding in a variety of ways depending on environmental factors. Two of the key factors that affect pedestrian movement as identified in previous literature are the spatial characteristics of a given setting, e.g., visibility, layout, or diversity, and the wayfinding support system, such as signage and information boards (Weisman, 1981).

Researchers aiming to observe wayfinding behaviour have developed operational models of individual behaviour using a symbolic artificial intelligence (AI) approach (Dridi, 2015). The goal in these models is to simulate human decision-making and problem-solving processes. An AI model's detailed behaviour though is not explicitly described by its algorithm, rather its behaviour is influenced by the specific problem presented to it (Moore, 2017). Although these new types of approaches, are employed in the study of pedestrian movement, evolving prediction methodologies, they fall short as a comprehensive theory of environmental psychology (Aschwanden et al., 2019, Van Dijk, 2018, Ye et al., 2019). Therefore, additional sources of information are an inevitable requirement of successful modelling in the field of pedestrian walking behaviour (Angelelli et al., 2018).

Space syntax

The Social Logic of Space by Hillier and Hanson (Hillier and Hanson, 1984) investigated historic cities and discovered that their organic development resulted in remarkably similar street patterns. They developed '*space syntax';* a set of techniques for describing and analysing spatial configurations in the context of human socioeconomics. Although the complex question of societal behaviour and spatial production was developed in the fields of architecture and urban planning in early 80s, leading to the investigation of various spatial models, such as space syntax research, one of its disadvantages is its limitation to measuring space in a static way (e.g., geometry or topology) (Hillier and Hanson, 1984; Till, 2007). Majority of these studies researching urban design and walkability rely on the assumption that street configuration is the most important influencing factor of pedestrian movement (Wang and Huang, 2019; Hillier and Hanson, 1984). A wide variety of researchers utilise configuration analysis to better estimate pedestrian flow volumes and route choices (Capitanio, 2019; Mansouri and Ujang, 2017; Özer and Kubat, 2015). This approach translates complex street networks into behavioural principles of the individuals' preference for high street network legibility (Boumezoued et al., 2020). Nevertheless, this lacks qualitative information, such as aesthetics of chosen routes, safety feelings, light conditions, and others, and impacts due to increased time spent in area or change in route directions.

Such simulation approaches do not reflect well the theory of environmental psychology. Although they present advantages relating to the high controllability of the study, such approaches require manual recording of information to verify simulation results, and although there are opportunities to capture unbiased behavioural data, this can result in relatively small sample sizes, not being representative of the population, and further lacking temporal information (Feng et al., 2021). In addition, movements of individuals are flexible, hence they cannot be considered as continuous over space, as they are entitled to the freedom of revisiting places or changing their movement decisions continuously in time, adding into the modelling exercise complexity. Therefore, the outcomes of these approaches are hypothetical and do not always represent reality as they can be biased and lacking in accuracy.

Research using novel data sources and methodologies

The increasing availability of data sources offered opportunities to researchers to renew the concepts and methods currently used in urban space design. A wide variety of new methodologies, often referred to as “Big Data Approaches” (BDAs), have become apparent including ML, network analysis and visualisation techniques, used to better capture and analyse a range of complex urban space problems (Aschwanden et al., 2019; Kontokosta et al., 2018; Díaz-Álvarez et al., 2018). BDAs have been employed to manage increasing urban data complexity in cities, with ML techniques employed in transportation and environmental studies (Aschwanden et al., 2019; Kontokosta et al., 2018; Díaz-Álvarez et al., 2018; Yue et al., 2022; Ali et al., 2021; Pollard et al., 2018). Unsupervised ML, and more specifically, clustering algorithms have been used in retail to reveal customer behaviours, while other researchers have used similar principles to cluster transit information based on temporal and spatial characteristics (Mauri, 2003; Chang and Chen, 2009; Ma et al., 2013).

Introduction of new sources and types of data therefore promise opportunities to better understand end-user needs, capturing ‘panoptic’ data which is not easy to observe in the real world and addressing problems at both the city and neighbourhood scale, whilst reducing traditional approach limitations, such as cost and scale implications from traditional data collection techniques. For example, employment of large-scale monitoring via smartphones and sensor networks (Wirz et al., 2013) enables researchers to study crowds in large settings. Key to this is the mobile phone, which has been used widely in urban planning and transportation sector applications (Shi and Abdel-Aty, 2015; Moreira and Ferreira, 2017; Martín et al., 2019), phones now being an integral part of human life. The device itself is transformed into a complex gadget that includes multimedia technologies that can reveal user preferences in terms of commercialism, daily routines, and cultural choices (Lee, 2011). These devices also include Wi-Fi and Bluetooth radio communication and connections of these to triangulated base station nodes can be used for precise geolocation.

Various studies, from indoor environments to transportation hubs and mass events, have applied Wi-Fi tracking techniques, video footage or traffic cameras to collect large-scale movements or to achieve real-time crowd monitoring (Duives et al., 2020; Peftitsi et al., 2020). For example, Duives et al. (Duives et al., 2020) combined video systems and computer vision algorithms to study pedestrian movement in mass events, while Li et al. (Li et al., 2020) used process imaging techniques to analyse pedestrian behaviour in zigzag corridors in the context of safety. Sensor based approaches, such as Wi-Fi tracking, present several advantages compared to other data sources, such as closed-circuit television (CCTV) video, Bluetooth, ultra-wideband transceivers (UWB) and others. The use of the Wi-Fi location tracking is very promising compared to other solutions and it represents a suitable solution for the following reasons: Cost effective solution; reasonable bandwidth, which leads to a high range resolution; wide coverage as Wi-Fi networks are extensively used in commercial, public and private sectors, overcoming some of the limitations of other approaches (e.g., cameras record a small area which is then difficult to stitch together or UWB can support specific types of devices); and reasonable transmitted power, which gives the Wi-Fi signal an advantage over short-range sensing technology such as UWB (Colone et al., 2011; Wang et al., 2018). However, a major disadvantage of a Wi-Fi location system is the positioning accuracy, as it depends on the experiment setup, hence it should be considered as a solution in relation to the research question.

The analysis of large amounts of data from large-scale wireless networks though can be challenging due to the inherent characteristics of the wireless environment, such as user mobility, noise, and data redundancy (Koh et al., 2020; Medeiros et al., 2020). In addition, although such techniques offer solutions to the manual collection of information compared to observational studies, there are many restrictions related to installation permissions in the public domain and ethical considerations. Such techniques offer opportunities for studying population patterns in larger contexts, while as pedestrians have limited knowledge of being tracked, results have a high degree of validity. However, the factors influencing pedestrian behaviour cannot be controlled, and the conditions under the data were recorded are not controllable by the researcher (Feng et al., 2021).

Study area

A high-street area in London, UK was selected as a case study site for the assessment of pedestrian walking patterns (Fig. 1). More specifically, Oxford Circus in London is a road junction and one of the busiest pedestrian crossings in the city, connecting two of the most prominent retail streets, Oxford Street and Regent Street, located in the London's West End. Oxford Street is a key

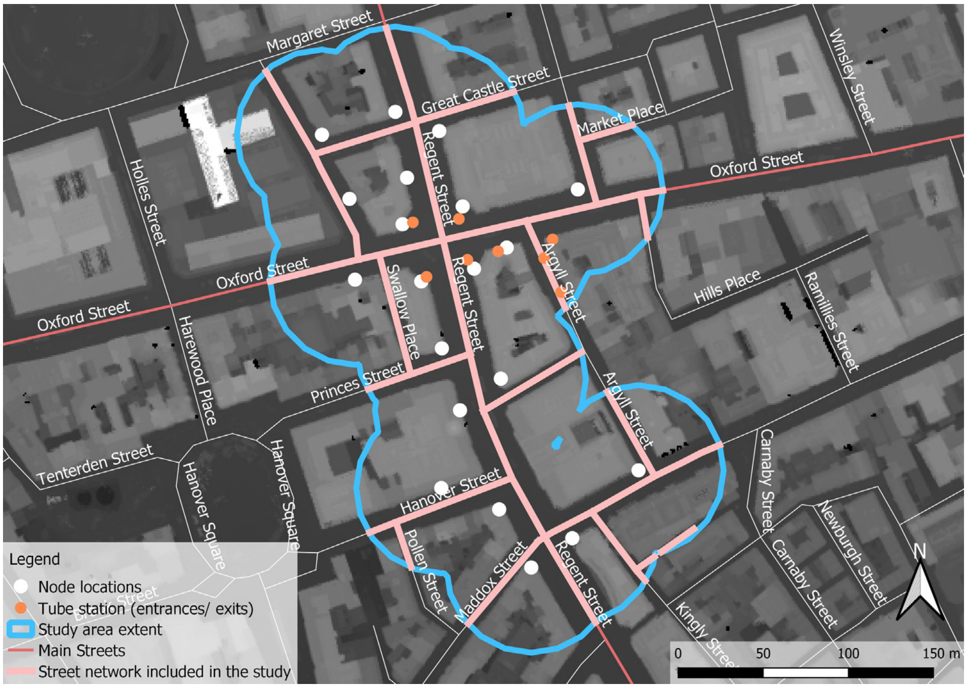

Fig. 1. Study area and the distribution of Wi-Fi nodes.

transportation corridor, used extensively by taxis and cyclists and providing east-west routes for bus services and tube stations. The Oxford Street district includes residential and retail areas, with commercial/office use. Regent Street is a major shopping street, containing flagship retail stores. Regent Street is c.1.3km long, while Oxford Street is 1.9km. Their intersection was transformed in 2011 from a segregated junction with barriers with limited overflow of pedestrians to an open diagonal crossing allowing pedestrians to follow their desired route. This change reflects a shift in street design towards the concept of integration and space sharing as a way of improving quality of environments, further enhanced by removal of street furniture, and with as many shared “single” surfaces as possible (Mercieca et al., 2011).

The study area location was chosen based on the availability of datasets, the mix of uses, and street networks connecting to wider residential areas. This therefore comprises an urban context with significantly different conditions and potentials, while it presents environments similar to those found in many other cities and urban areas (Carmona, 2015). The study area extent is defined by the location of the Wi-Fi nodes and their coverage. This includes all the streets located within the coverage area of the nodes (Blue line in Fig. 1). This is further explained in the following sections.

Materials and methods

Data preparation

This study utilises multiple data sources, allowing a number of spatial attributes to be explored. These data include pedestrian movement, urban geometry, and weather information data. All datasets selected, excepting the Wi-Fi location data, are publicly available data, described in Table 2. Data information reflect the category of information captured in each dataset, while geometry type and date describe the shape of the captured information and the timeframe in which it was captured.

Pedestrian movement data, obtained by Wi-Fi tracking, was provided by The Crown Estate (The Crown Estate 2021), with data deriving from an earlier study aiming to inform a base year model used to simulate existing pedestrian flow conditions and predictive impact of changes in demand or spatial layouts (Angelelli et al., 2018). Accuracy of the data has been tested in that same study, based on comparison of the Wi-Fi obtained data and CCTV image data. Data was collected by capturing signals from Wi-Fi devices across 19 Open Mesh nodes (OM2P-HS) attached to floodlights on building cornices (Fig. 1 – node location). The nodes were installed at 3 to 5 m height, to be clear from obstructions that could affect signal reception. Data was transmitted in real-time via the 3G/4G mobile network for storage in a cloud-based server.

The mobile data used covers the period August to October 2017 (Fig. 2), in total 22 days, incorporating 3,240,361 unique mobile users. The August period presents the biggest continuous data sample and is the only period including weekend data collection. Data was pre-processed by the technology provider (Accuware Inc 2017) eliminating all privacy-related information, providing outputs as a multi-field .csv file per day, capturing unique Media Access Control (MAC) addresses, signal strengths, location in X Y Z coordinates and a timestamp. MAC addresses are unique identifiers assigned to a network interface controller for use as a unique network address,

Table 2
Types of data collected for this study with source.

Data information	Geometry type	Date	Source
Pedestrian Data	Point	2017	Wi-Fi tracking
Important buildings & Infrastructure	Point	2018	Ordnance Survey (https://osmaps.ordnancesurvey.co.uk/) Open street map (https://www.geofabrik.de/geofabrik/) POIS (https://osmaps.ordnancesurvey.co.uk/) Registered EPC non-domestic (www.gov.uk)
Park areas	Polygon	2018	Ordnance Survey (https://osmaps.ordnancesurvey.co.uk/) Open street map (https://www.geofabrik.de/geofabrik/)POIS (https://osmaps.ordnancesurvey.co.uk/)Registered EPC non-domestic (www.gov.uk)
Transportation access (bus stops/ tube entrances) location	Point	2018	Ordnance Survey (https://osmaps.ordnancesurvey.co.uk/) Open street map (https://www.geofabrik.de/geofabrik/)POIS (https://osmaps.ordnancesurvey.co.uk/)Registered EPC non-domestic (www.gov.uk)
Street geometry	Polygon	2018	Ordnance Survey (https://osmaps.ordnancesurvey.co.uk/) Open street map (https://www.geofabrik.de/geofabrik/)POIS (https://osmaps.ordnancesurvey.co.uk/)Registered EPC non-domestic (www.gov.uk)
Amenities	Point	2018	Ordnance Survey (https://osmaps.ordnancesurvey.co.uk/) Open street map (https://www.geofabrik.de/geofabrik/)POIS (https://osmaps.ordnancesurvey.co.uk/)Registered EPC non-domestic (www.gov.uk)
Hourly temperature, humidity & weather events	Point (temporal)	2017	Weather Underground/ Private weather stations

common in technologies, such as Ethernet, Wi-Fi, or Bluetooth. Signal strength was used for triangulation purposes by the provider to derive location, indicating nodes in greater proximity to the recorded devices. Data were being captured on a frequency varying from 1-60 seconds, dependent upon the type of handset devices, manufacturer, or activity level. The Wi-Fi tracking system used in this study recorded signals with an accuracy of -3/+3m.The Wi-Fi nodes can capture the presence of a device at a distance of up to 100 m, while the signal strength from devices decreases exponentially with distance (Accuware Inc 2017).

All the data handling, storage, processing, and presentation observe the data security and privacy requirements as specified in General Data Protection Regulation (GDPR) on handling personal data and the protection of privacy (EUROPEAN PARLIAMENT AND OF THE COUNCIL 2016). Personal information was truncated via the system and converted to non-personal form, thus permitting collection of information without consent (Fuxjaeger and Ruehrup, 22 February 2018).

Conceptual model

Location data were assessed using a data analysis conceptual model developed based on the data collected (Fig. 3). Analysis was performed on a daily resolution utilising the following steps:

- Data pre-processing: Data pre-processing utilising bespoke algorithms and space syntax methodologies was used to extract valuable information and to enrich existing datasets. Walking pedestrian characteristics, spatial visibility and weather information were mapped against individual points recorded via Wi-Fi tracking technique, described in detail in the following sections.
- Data preparation for K-means clustering analysis and model development.
- Data cleaning and normalisation using outlier removal (Interquartile Range Method) (Vinutha et al., 2018) and Min-Max scaler method (Han et al., 2012).
- Multiple factor analysis to remove multi-dimensionality of the data and to select key variables for the model.
- Analysis & Results: Cluster analysis (unsupervised machine learning) to extract key behavioural patterns and identify classes of homogeneous profiles.

Finally, based on the similarities returned by the clustering analysis, a data mapping exercise against these categories was undertaken to investigate three key hypotheses: (i) recorded speed and purpose present a clear and consistent relationship, (ii) visibility as a driver for movement, and (iii) knowledge of the area based on number of unique recorded devices, representing repeated visits. The level of experience within the area was calculated by identifying and counting instances with only one visit recorded through the number of days of the recorded dataset across all the days, with the assumption that this implied non-regular use of the area.

Data pre - processing

Following the conceptual model development, a bespoke algorithm written in Python was used to extract the variables calculated on a point-by-point basis. The point-by-point method follows the pattern of subtracting the n+1 point from the n point to extract the absolute values, where n is the first recorded location of the device within the study area. The variables extracted are: (i) Duration in seconds, (ii) Distance in metres, (iii) Speed in m/s, (iv) Bearing in degrees and (v) Day period (Table 3). To remove false recordings, a threshold for outlier distances was set at 5,000m, removing all points with such recorded distances. Bearing was calculated to provide

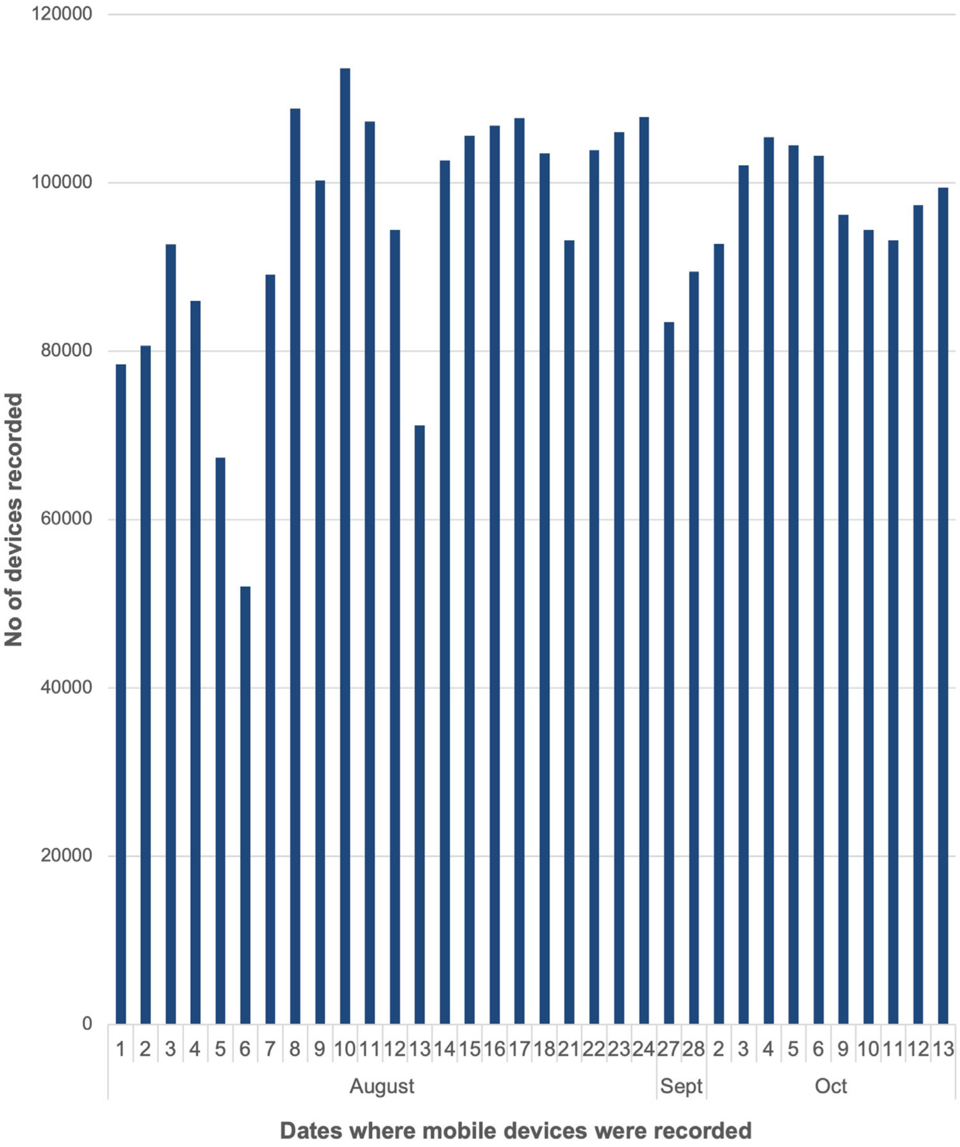

Fig. 2. Sample size: number of devices recorded for the 22 days (all data per day breakdown).

an indication of direction. The transformation to geographical degrees considers the geographical north at 0/360°, the Python code thus calculating a bearing as N=0/ 360; E=90; S=180; W=270. Day period classification was undertaken as follows, based on August 1st sun cycle in London, U.K. which for consistency it was used for all the study days:

- 'First light': 05:24 to 08:40
- 'Morning': 08:40 to 12:10
- 'Lunchtime': 12:10 to 14:00
- 'Afternoon': 14:00 to 20:47
- 'Last light': 20:47 to 21:28
- 'Nighttime': 21:28 to 05:24

A key limitation with the use of such big datasets is that the processing of information required polynomial run time and the researchers had to utilise the High-Performance Computer (HPC) facility of Cranfield University to overcome this issue (Cranfield University 16 August 2017). Sixteen CPU cores were needed, with a three-hour simulation time required per input csv file.

To better understand pedestrian movement in relation to spatial attributes, further information was acquired, using the space syntax methodology and more specifically, the visibility graph analysis (VGA), serving as a constant for spatial visibility (Turner et al.,

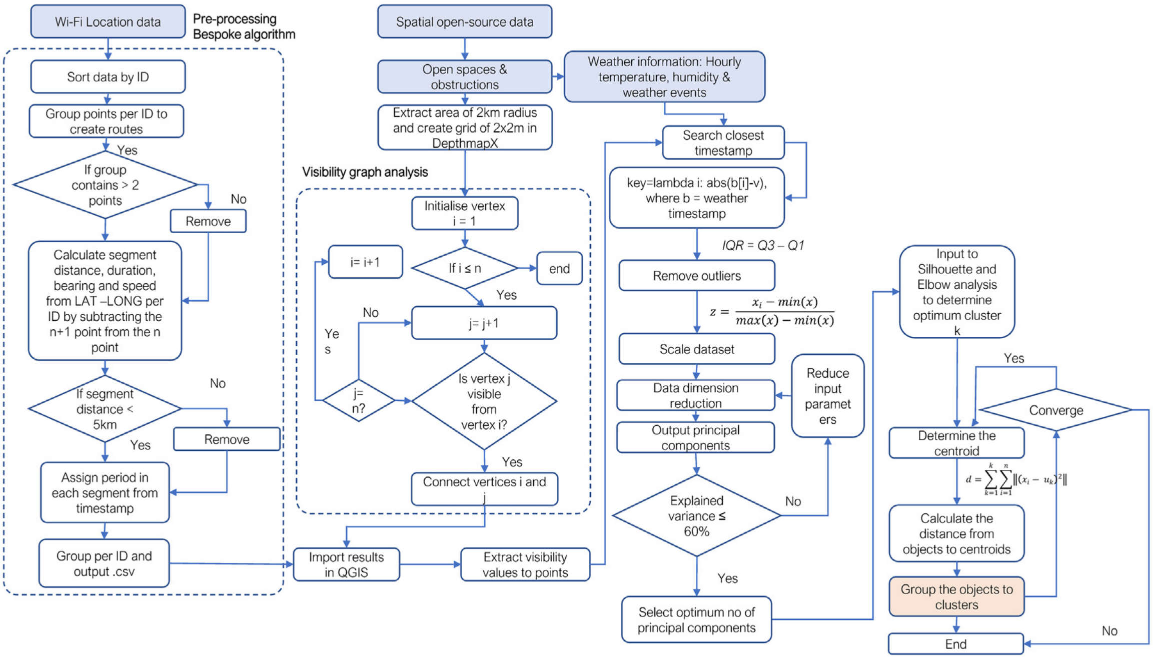
Wi-Fi Location data
Pre-processing Bespoke algorithm
Sort data by ID
Group points per ID to create routes
If group contains > 2 points
Yes
No
Remove
Calculate segment distance, duration, bearing and speed from LAT –LONG per ID by subtracting the n+1 point from the n point
If segment distance < 5km
No
Yes
Remove
Assign period in each segment from timestamp
Group per ID and output .csv
Spatial open-source data
Open spaces & obstructions
Weather information: Hourly temperature, humidity & weather events
Extract area of 2km radius and create grid of 2x2m in DepthmapX
Visibility graph analysis
Initialise vertex i = 1
i= i+1
If i ≤ n
end
Yes
No
Ye s
j= j+1
j= n?
Is vertex j visible from vertex i?
Yes
Connect vertices i and j
Import results in QGIS
Extract visibility values to points
Search closest timestamp
key=lambda i: abs(b[i]-v), where b = weather timestamp
IQR = Q3 – Q1
Remove outliers
z = (x_i − min(x)) / (max(x) − min(x))
Scale dataset
Data dimension reduction
Reduce input parameters
Output principal components
Explained variance ≤ 60%
No
Yes
Select optimum no of principal components
Input to Silhouette and Elbow analysis to determine optimum cluster k
Yes
Converge
Determine the centroid
d = Σ_{k=1}^{k} Σ_{i=1}^{n} ‖(x_i − u_k)^2‖
Calculate the distance from objects to centroids
Group the objects to clusters
No
End

Fig. 3. Overview of proposed algorithm flow of the conceptual model (blue boxes represent inputs and orange box represents outcome.

Table 3
Variable set used.

Variable name	Data type	Description	Year collected	Source
ID	Categorical	MAC address	2017	Wi-Fi tracking
end_lon	float	X coordinate	2017	Wi-Fi tracking
end_lat	float	Y coordinate	2017	Wi-Fi tracking
Period	Categorical	Name of the assigned period	2017	Wi-Fi tracking
end_time	Categorical	Date and time	2017	Wi-Fi tracking
bearing_segment	float	Bearing in degrees	2017	Wi-Fi tracking
duration_segment	Integer	Time spent in seconds	2017	Wi-Fi tracking
distance_segment	float	Distance travelled in metres	2017	Wi-Fi tracking
speed_segment	float	Walking speed in m/s	2017	Wi-Fi tracking
rvalue_1	float	Spatial visibility	2018 (Building polygons used for VGA simulation)	https://osmaps.ordnancesurvey.co.uk/, https://www.openstreetmap.org, https://www.geofabrik.de/geofabrik/)
Humidity_%	float	Humidity in percentages	2017	Weather Underground. Weather Station ID: ILONDON636
Speed_mph	float	Wind speed in mph	2017	Weather Underground. Weather Station ID: ILONDON636
Precip. Rate _in	float	Precipitation rate in inches	2017	Weather Underground. Weather Station ID: ILONDON636
Solar_w/m^2	float	Solar radiation in w.m^2	2017	Weather Underground. Weather Station ID: ILONDON636
Temperature_C	float	Temperature in Celsius	2017	Weather Underground. Weather Station ID: ILONDON636
hours	float	Hour of the day	2017	Wi-Fi tracking

2001). For the purposes of this paper, space syntax theories were employed using VGA assessments via the open-source software DepthmapX_net_035 (Varoudis, 2021). An area of 2km diameter was adopted as the distance threshold, preventing result distortion due to the small scale. According to Ahrné et al. (2009), minimum distance thresholds range from 300m to 1km radius, hence the 2km diameter scale was chosen by the authors (Ahrné et al., 2009). Following the VGA, the use of a Geospatial Information Systems (GIS) platform and the function “*Extract Values to Points”* was used, and values were mapped against each point, serving as an additional analysis parameter.

Finally, weather data were exported from a private weather station located in the area for each date. Weather information were recorded every five minutes and the most appropriate fields were selected to reflect the microclimate conditions in the study area. The parameters comprised humidity, wind speed, precipitation, solar radiation, and temperature. All weather information was mapped against the location dataset, using the bisect method (array bisection algorithm) in Python. The full set of variable inputs compiled are displayed in Table 3.

Choosing ML clustering method: k-means

Clustering algorithms are classified into several types based on their partitioning, density, and model (Zhai et al., 2014). A clustering algorithm divides a physical or abstract object into a group of related things (Yuan and Yang, 2019). A partition-based clustering algorithm is required in this study as the goal was to exclusively segregate the input observations so that each point belongs to one group only. The K-means algorithm has numerous advantages in comparison to other recognised methods, including simple mathematical concepts, rapid convergence, better scaling to large datasets, efficient handling of high dimensional datasets and ease of implementation (Li et al., 2017). Additionally, this method can be applied in a broad range of fields, and it can be easily adapted to new examples.

Data preparation for K-means analysis and model development

Cluster analysis was utilised to reveal walking behaviours and to identify key groups in the case study area. ML pattern-mining techniques are generally used to identify unknown patterns within normalised datasets (Abu-Bakar et al., 2021). Cluster analysis labels observations (data points) within assigned groups, or ‘clusters’, extracting key patterns and identifying classes of homogeneous profiles. K-means algorithms partition data into clusters by minimising the within-clusters sum-of-squares (Yuan and Yang, 2019, Wang et al., 2012) (Eq. 1).

$$d = \sum_{k=1}^{k} \sum_{i=1}^{n} \left(x_i - u_k\right)^2 \tag{1}$$

where d is the main function of sum of the squared error, k is the number of clusters, n is the number of observations, x_i is observation i and u_κ is the centroid formed for x_i’s cluster. The mean of the recorded data is constantly updated, and each observation is placed within the cluster having the nearest centre until no more observations can be assigned (Forgy, 1965).

This method was chosen due to the nature of the input observations and the manner by which this method exclusively segregates clusters, so as each point belongs to one group only, where each partition is represented by one cluster only and k ≤ n (Han et al., 2012; Zhu et al., 2010). As traditional K-means also present several limitations, improvements on the method and data preparation have been suggested by previous research to receive better results when solving practical problems (Wagstaff et al., 2001; Huang, 1998; Narayanan et al., 2016; Narayanan et al., 2019). To overcome such limitations and building upon previous literature, the following steps were performed to ensure data suitability for the unsupervised ML model (Celebi et al., 2013; Zhang and Leung, 2003; Namratha Reddy and Supreethi, 2017), namely: (1) Input variables limited to numerical only, (2) Noise & outlier removal, (3) Data normalisation, (4) Reduction of the number of variables, (5) Collinearity, (6) Determining the optimal number of clusters. Each is discussed below.

1) Input variables limited to numerical only

K-means uses distance-based measurements to determine similarity between data points, therefore numerical variables are the only input that can be processed. Additionally, removal of undefined (NaN) values was performed following the first .csv output from the raw data analysis and the weather mapping values. NaN values cannot be considered as 0, as the 0 value is a meaningful in this type of analysis. Nevertheless, the NaN values were less than 10%, an acceptable percentage when dealing with missing values (Bennett, 2001). Listwise deletion (LD) was used to remove all the NaN values (Peng et al., 2006).

2) Noise & outlier removal

K-means is sensitive to outliers and ‘noisy’ data (Jin and Han, 2011). If data is not pre-processed to remove noise and outliers, then K-means can return false results, driven by the strongest set of information. Outlier values were therefore removed using the interquartile range method for each individual variable (Vinutha et al., 2018).

3) Data normalisation

For the ML algorithm to consider all attributes as equal, they must all have the same scale, hence the Min-Max Scaler method was used, implemented via Python and the help of scikit-learn package (Han et al., 2012; Thorndike, 1953). This method was chosen as it transforms each value in the columns proportionally, within the bounded intervals, achieving a linear transformation on the original data (Abu-Bakar et al., 2021). The Min-Max Scaler method is considered ideal for revealing patterns by highlighting any peaks or falls in a consistent manner (Abu-Bakar et al., 2021).

Each variable was then transformed by scaling to the range 0-1 (Eq. 2).

$$z = \frac{x_i - min(x)}{\max(x) - min(x)} \tag{2}$$

where z is the normalised value, x_i is the original value range, min(x) is the minimum range attribute and max(x) is the maximum attribute range. This step ensured that different scales would not skew the results and would contribute equally to model fitting.

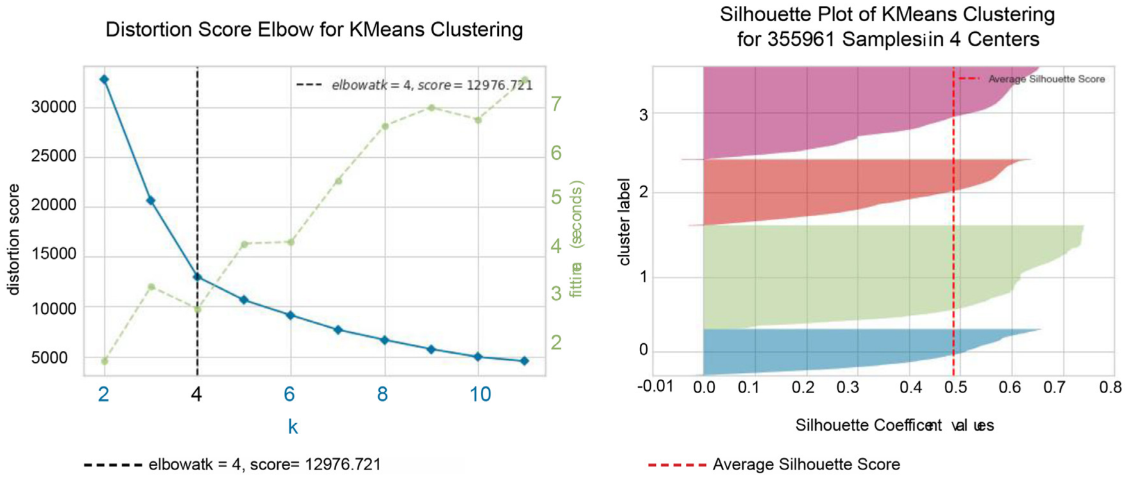

Fig. 4. Elbow (left) and silhouette analysis (right) results example on a typical day in August (5th).

4) Reduction number of variables

As the number of variables increases, a distance-based similarity measure converges to a constant value between any given points. The more variables, the more difficult to find strict differences between instances. One of the most popular approaches to dimensionality reduction is principal component analysis (PCA). This method seeks to reduce the number of variables whilst preserving the most important structure and relationships of the input data, while it has also been shown to produce improved results when dimensionality of datasets is high by comparison with other methods (Reddy, 2020).

To define the number of variables, the explained variance ratio metric was used for the first two components. The explained variance ratio is the percentage of variance attributed by each of the selected components. Ideally, the number of components chosen for model inclusion is decided by adding the explained variance ratio of each component until a total of around 0.8 or 80% is attained to avoid overfitting (Joliffe and Morgan, 1992). The variance should not be less than 60% (Hair et al., 2014). Where the variance explained is 35%, it indicates the data is not useful and may need further measures. If the variance is less than 60%, there are most likely chances of more variables that can be apparent. The explained variance ratio returned with all the variables, excluding speed, was 45%. Therefore, additional parameters were removed, resulting in selection of only the variables of (1) visibility, (2) bearing, (3) distance, (4) duration (Table 2); the explained variance ratio was then recalculated as 76%.

5) Collinearity problem

Correlated variables are not useful for ML segmentation algorithms, representing the same characteristic of a segment (e.g., noise). Speed was tested with PCA analysis, as a suitable variable, but was removed from the final list of input variables due to a high level of correlation with duration and distance (Reddy, 2020).

6) Determining the optimal number of clusters

Clustering algorithms rely on a random initialisation of the cluster centre. To overcome this issue, the Elbow Method (EM) (Thorndike, 1953) and Silhouette Analysis (SA) (Rousseeuw, 1987) were undertaken to reveal the ideal number of clusters, where a randomised seeding technique guarantees the optimal solution is obtained. The EM is one of the most popular methods used to determine the optimal value of k, using two calculation metrics: distortion and inertia (Han et al., 2012). Distortion is the average of the squared distances from the cluster centers of the respective clusters, and typically the Euclidean distance metric is used. Inertia is the sum of squared distances of samples to their closest cluster center. K-means clusters data by separating samples to n groups of equal variances, minimizing the criterion known as the inertia or within-cluster sum-of-squares (WCSS). Therefore, the smaller the inertia the denser the cluster (the closer the points are). The Silhouette Score ranges from -1 to 1 indicating how close or distant the clusters are from each other and how dense the clusters are.

Analysis and results

The analysis from the EM and SA method application identified the number of clusters within the data. The optimum number of clusters was returned as k=4 (Fig. 4), indicating there are four different groups of behaviours. This test was performed initially using the daily resolution; however, it was then performed again in both period and hourly resolutions, all returning same results. Examples of the returned values are shown in the graphs below for the 5th of August 2017 (Fig. 4). The analysis was performed in Python with the help of scikit-learn package (Scikit-learn 0.19.1 documentation 2018).

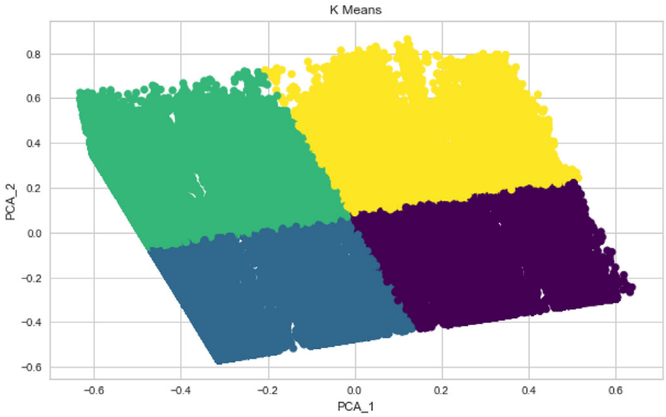

Fig. 5. Cluster results example (5th of August 2017). Each colour represents a different cluster.

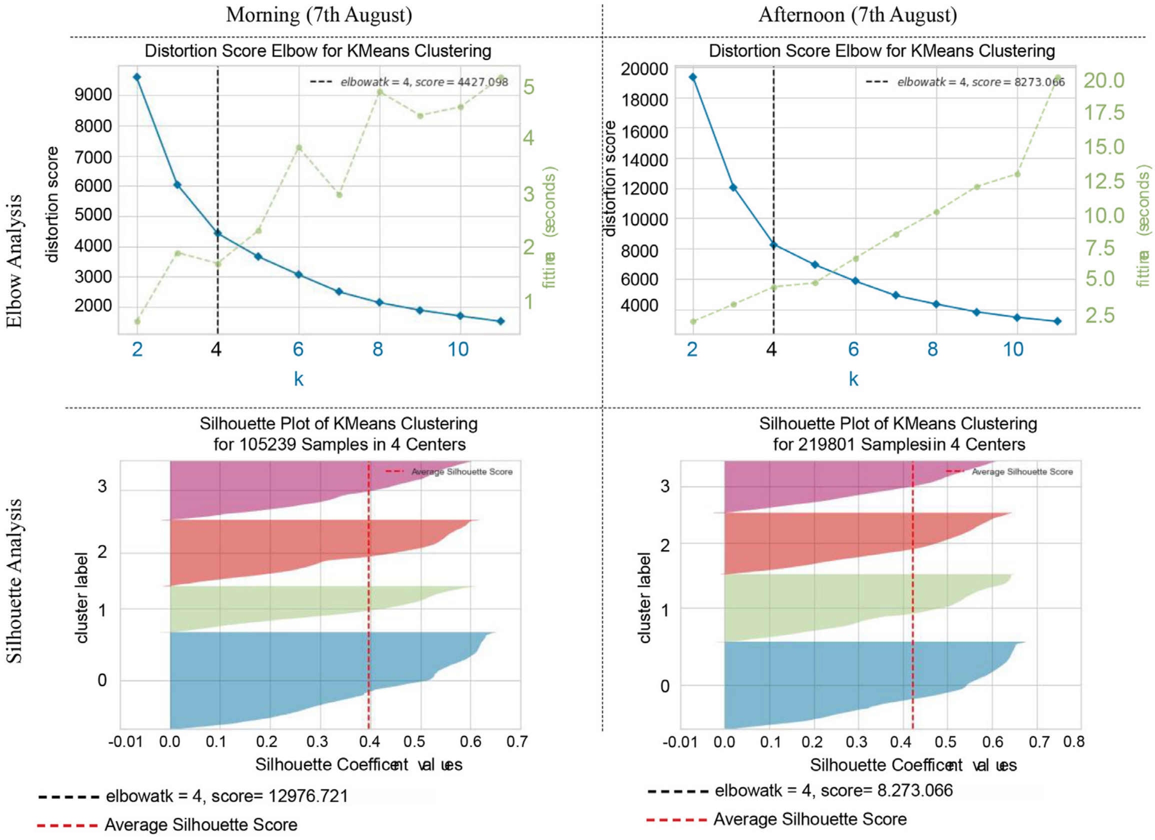

Fig. 6. Analysis results in period resolution in a typical weekday (EM and SA methods).

Cluster analysis was performed with k=4 (maximum iteration = 100 and random state = none) as indicated by the EM and SA, for each day of the overall dataset, due to the difference in patterns and behaviours occurring in different times of the year and weekdays. The input data were classified using the clustering algorithm and organised into classes sharing similar attributes (Fig. 5).

The clustering steps were repeated for key days to validate previous analysis. Results revealed that the same number of clusters exist, via the use of EM, followed by acceptable values in SA (Fig. 6).

An additional metric was selected to further validate the number of optimum clusters, the Calinski-Harabasz coefficient (CH), and analysis for this was undertaken for both resolutions (daily & period) (Fig. 6). The CH, also known as the variance ratio criterion, is a measure based on the internal dispersion of clusters and the dispersion between clusters. CH criterion was chosen as a validation method as it is fast to compute, appropriate due to the great amounts of data in this study and is established as one of the best performing methods for estimating numbers of clusters (Milligan and Cooper, 1985). In addition, this metric is suitable when dealing

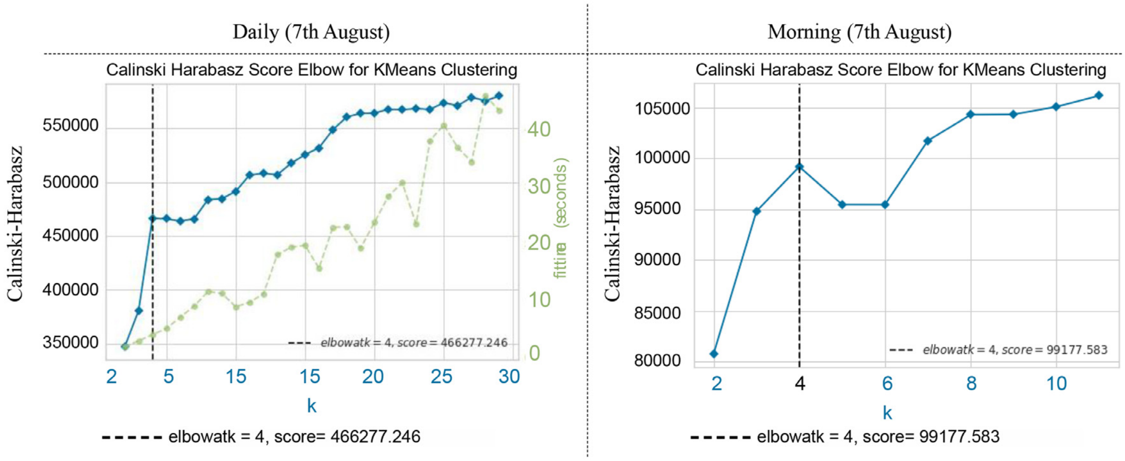

Fig. 7. Analysis results examples in daily & period (morning) resolution (Calinski-Harabasz coefficient).

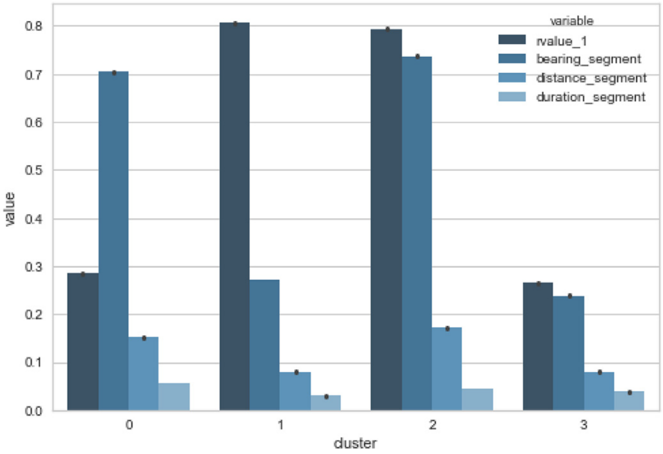

Fig. 8. Point-based feature extraction per cluster and importance ranking of variables (example 7th August).

with compact clusters (convex) (Calinski and Harabasz, 1974). The optimum number of clusters is the number that maximises the CH value (Calinski and Harabasz, 1974), without overfitting the model. The optimum number of clusters is chosen when a peak or an elbow on the line plot of CH indices. The analysis was performed in Python with the help of scikit-learn package, returning the optimum number of clusters as k= 4 (Fig. 7).

Following cluster analysis, feature importance extraction was performed in Python to better understand clustering results drivers, where each feature receives a score indicating that the higher the value, the more important or relevant it is towards the output variable. The results revealed that the main driver order for the clustering results was as follows: (i) spatial visibility (rvalue_1), (ii) route direction (bearing_segment), (iii) distance travelled (distance_segment), and (iv) time spent in point (duration_segment). Fig. 8 shows ranking importance displayed in the correct order in the graph's legend, named "variable".

Clustering results revealed four distinct clusters for all the individual days assessed. Extraction of the key characteristics for each cluster was undertaken and a summary of results identified (Table 4). A descriptive summary of the results included count, mean standard deviation, minimum and maximum values, plus lower and upper percentiles and the 50th percentile (median). Counting revealed cluster results to be balanced, with the sample sizes similar in all categories (Amin and et al., 2016). Results illustrate that Cluster 0 and 2 users are generally travelling longer distances, with the mean distance being 14.06 and 12.16m respectively between recorded points, whereas Clusters 1 and 3 users travelled 5.78 and 7.11m. Clusters 0 and 3 spent time in areas of high visibility, with mean rvalue_1 recording being 13,784.69 and 13,988.06 respectively. The shortest time spent at each point segment was observed in Cluster 1, with a mean value of 31.15 seconds, while the highest was in cluster 0, with a mean value of 66.74 seconds.

Hypothesis 1: recorded speed and purpose

Cluster 0 and 2 have the highest recorded walking speeds in the August period, while Cluster 1 and 3 have the lowest (Fig. 9 – showing walking speed heatmaps), indicating that at this time, pedestrians' goals for Cluster 0 and 2 users were better defined than

Table 4
Descriptive statistics of the clustering results (August 2nd to 18th).

Cluster 0

	bearing_segment	duration_segment	distance_segment	speed_segment	rvalue_1	cluster
count	1,932,972.00	1,932,972.00	1,932,972.00	1,932,972.00	1,932,972.00	1,932,972.00
mean	235.55	66.74	14.09	0.40	13,784.69	0.00
std	54.93	146.15	19.41	0.52	6,762.04	0.00
min	76.09	5.00	0.00	0.00	136.15	0.00
25%ile	198.43	5.00	1.00	0.06	7,993.95	0.00
50%ile	216.03	15.00	4.00	0.18	14,710.07	0.00
75%ile	286.80	55.00	21.00	0.52	18,345.94	0.00
max	341.57	1,370.00	87.00	2.69	33,442.28	0.00

Cluster 1

	bearing_segment	duration_segment	distance_segment	speed_segment	rvalue_1	cluster
count	2,547,052.00	2,547,052.00	2,547,052.00	2,547,052.00	2,547,052.00	2,547,052.00
mean	92.31	31.15	5.78	0.30	38,045.32	1.00
std	59.28	92.78	11.42	0.44	6,843.82	0.00
min	0.00	5.00	0.00	0.00	22,783.48	1.00
25%ile	39.81	5.00	0.00	0.01	34,572.00	1.00
50%ile	90.00	10.00	1.00	0.15	36,654.67	1.00
75%ile	139.50	20.00	5.00	0.36	45,839.99	1.00
max	339.06	1,370.00	87.00	2.69	48,533.89	1.00

Cluster 2

	bearing_segment	duration_segment	distance_segment	speed_segment	rvalue_1	cluster
count	2,105,173.00	2,105,173.00	2,105,173.00	2,105,173.00	2,105,173.00	2,105,173.00
mean	225.38	43.49	12.16	0.40	35,218.27	2.00
std	73.08	112.71	18.80	0.49	10,114.79	0.00
min	0.00	5.00	0.00	0.00	509.05	2.00
25%ile	183.37	5.00	1.00	0.09	29,216.61	2.00
50%ile	215.91	10.00	3.00	0.20	36,372.39	2.00
75%ile	287.65	30.00	15.00	0.51	44,098.32	2.00
max	341.57	1,370.00	87.00	2.69	48,533.89	2.00

Cluster 3

	bearing_segment	duration_segment	distance_segment	speed_segment	rvalue_1	cluster
count	2,726,997.00	2,726,997.00	2,726,997.00	2,726,997.00	2,726,997.00	2,726,997.00
mean	79.88	46.18	7.11	0.24	13,988.06	3.00
std	42.45	118.99	14.71	0.44	8,600.58	0.00
min	0.00	5.00	0.00	0.00	136.15	3.00
25%ile	35.54	5.00	0.00	0.00	7,403.70	3.00
50%ile	90.00	10.00	1.00	0.05	14,307.09	3.00
75%ile	98.39	25.00	5.00	0.25	18,145.74	3.00
Max	192.80	1,370.00	87.00	2.69	48,212.39	3.00

those in the other two clusters. The lowest speed was recorded for Cluster 3, with an average value of 0,24 m/s in August and 0,23 in October, while Cluster 2 has the highest speeds recorded in August with a mean value of 0,42 m/s. However, as this research explores walking patterns on a point-by-point basis rather than on a journey purpose basis, these types of classifications can only serve as baseline information.

Hypothesis 2: visibility as a driver for movement

Clusters 0 and 3 were revealed to have the lowest mean visibility values, indicating that at the specific points recorded, user movement was within areas with a limited space visibility (Fig. 10). Therefore, the results indicate Cluster 0 and 3 as having behaviours that appear when an individual has an idea of where places are or has an increased knowledge of the area, indicating that visibility is not a crucial parameter for their movement. Other needs, such as route efficiency, can be used to better define walking patterns.

Hypothesis 3: knowledge of the area based on number of unique recorded devices

Results indicated that Cluster 1 and 2 have the highest percentage of unique devices recorded, within the study period, from first to last light (05:24 to 21:28). This suggests that majority of the points recorded in these clusters are users with only limited knowledge of the area (Fig. 11). Fig. 11 shows the number of recorded devices for each cluster in a typical week in August. This graph indicates an increase in the number of unique users during weekends, while there is a considerable decline of unique users on a Monday (7th of

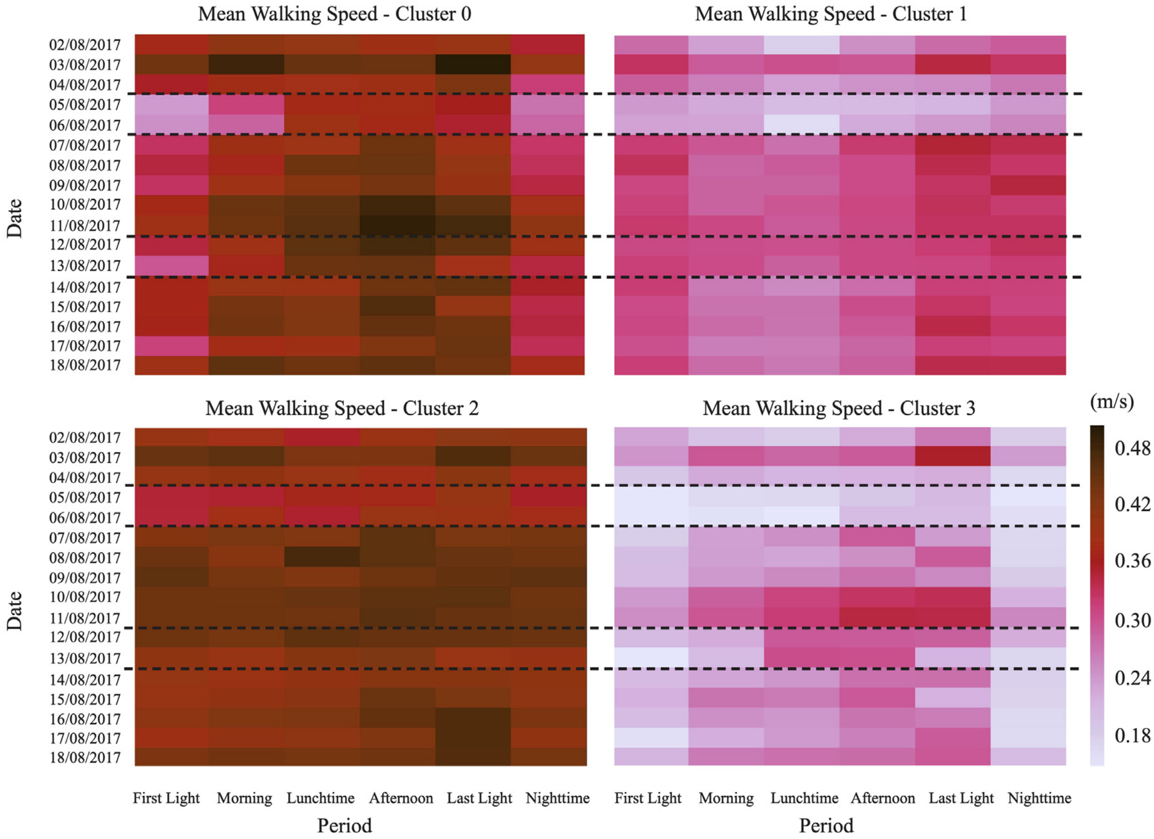

Fig. 9. Heatmaps illustrating mean walking speeds in m/s as recorded for each cluster in August. The cluster number is indicated on the top of each graph. Dates are shown on the vertical axis, for dates 2nd of August (top) until 18th of August (bottom), while period is indicated on the horizontal axis, dashed lines indicating weekends.

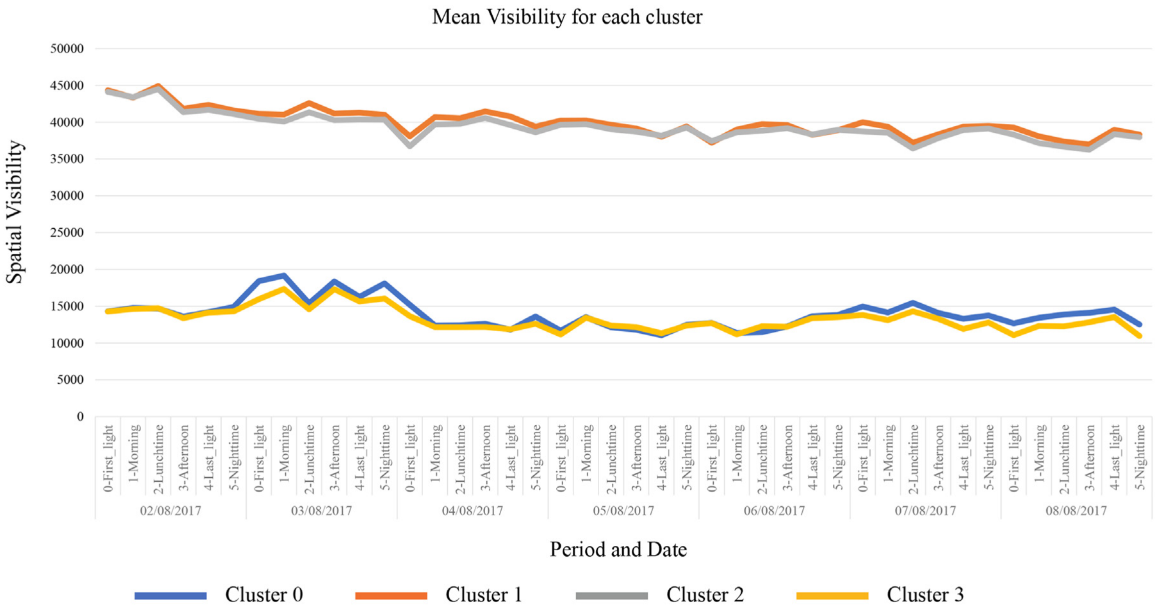

Fig. 10. Line graph illustrating mean space visibility for each cluster in a typical week in August (2nd to 8th). Mean values are grouped by period with colours representing clusters.

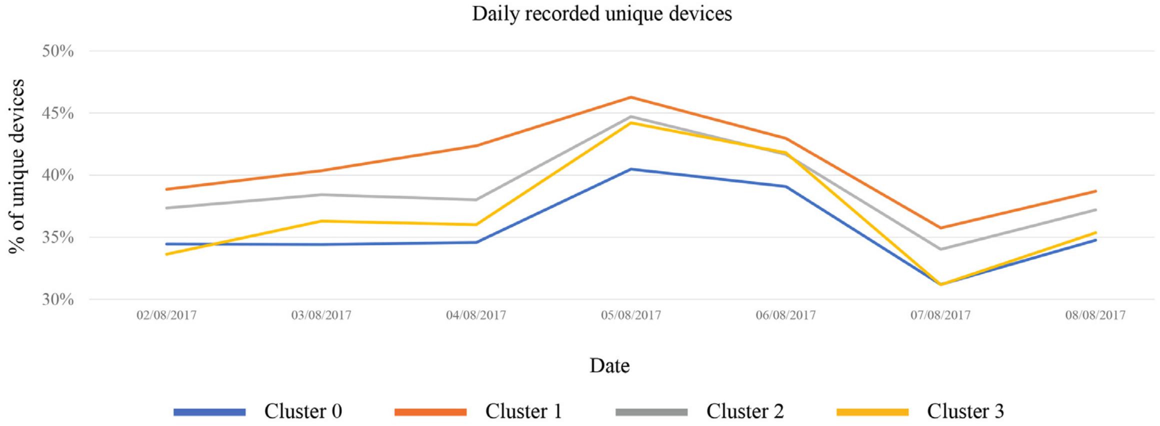

Fig. 11. Unique devices recorded for each cluster in a typical week in August. Values are daily and count devices from first to last light (05:24 to 21:28). Colours represent clusters. 5th and 6th of August are Saturday and Sunday respectively.

August, a UK Summer Day). Cluster 3 indicates that a greater majority of users visits the area during weekends, however, due to lack of weekend days (only 2 weekends were incorporated within August due to source data restrictions), these devices are highlighted as unique within the existing dataset.

Discussion

This paper broadens the understanding of the complex, multi-layered relationships occurring between urban space and pedestrians, via the combination of quantitative location data pertaining to end-users, as well as quantitative spatial attributes, such as visibility. This novel analytical approach provides a new interpretation of behaviours exhibited in an urban space while implying that although spatial configuration influences walking behaviour, such information crucially affects the quality of the walking experience and individual preferences. The 'Big Data' approach adopted further permits an automation to be applied to the analysis, making the approach suitable for large area analysis.

Although there are many different approaches to collect and analyse large scale data, there lacks a standardized framework to extract insights (Koh et al., 2020). Different contexts can reveal different behaviours, as urban environments vary significantly. In this paper, employment of an unsupervised ML conceptual model is proposed for identifying the pedestrians' intention while walking in urban space in a systematic way. The results illustrate how employment of unsupervised ML algorithms can reveal categories of pedestrian behaviours that can directly inform urban models, rather than depending on biased observational data or existing literature. Additionally, analysis of large-scale monitoring data in mobility tracking, such as via the use of Wi-Fi signals, presents limitations due to the lack of labelled data as these are passively collected without the subjects' knowledge (Harari et al., 2016; Medeiros et al., 2020). As a result, previous literature utilising ML algorithms and Wi-Fi tracking was focused on the quantitative aspects of movement patterns such as the fluctuation of people counts throughout the day, clusters of trajectories belonging to distinct groups of individuals, or relative flow volumes between different buildings (Koh et al., 2020; Mauri, 2003, Chang and et al., 2020; Ma et al., 2013), rather than the qualitative aspects of movement behaviour, such as experience. The results revealed that by mapping tracked pedestrian movement data to urban attributes, such as spatial visibility, enables the extraction of insights on pedestrian experiences and preferences to inform urban planning and design decisions.

It is suggested that the combination of the type of activity undertaken, and the knowledge of the area represent the key behaviours found in walking patterns in the context of a high-street, and that they can be grouped into two activity categories: (i) Utilitarian walking (with motivation/ destination) and (ii) Leisure walking (no motivation) (Ki and Lee, 2021). The first category includes learning and journey efficiency, where the user has specific destinations to reach with varying levels of area knowledge (expert and novice strider (Davies, 2007)). The second category includes wandering and open-ended journeys, where travellers explore already known areas of the city or discover them as they walk (expert and novice stroller (Davies, 2007)). Validation of the activities was undertaken utilising one hypothesis: (i) recorded speed and purpose. Mean daily walking speed recorded values in August range from 0,19 to 0,44 m/s. Cluster 0 and 2 had a mean value of 0,40 and 0,42 m/s respectively, indicating that the pedestrians were undertaking utilitarian activities, while Cluster 1 and 3 had lower mean values, 0,30 and 0,24 m/s respectively, indicating that these users fall within the leisure walking category (Ki and Lee, 2021). In addition, although literature suggests an estimated speed of 1,33 m/s (Bohannon, 1997), such findings indicate that either external parameters are causing pedestrians to slow down, or that a great majority of the recorded pedestrians are not moving.

Further to this, results indicated that a consistent link exists between spatial visibility and previous experience of the area. The greater the familiarity a pedestrian has with the surrounding space, the less spatial visibility is needed for its pedestrian movement (Fig. 12). Validation of knowledge of the area was undertaken utilising two additional hypotheses: (ii) visibility as a driver for movement and (iii) knowledge of the area based on number of unique recorded devices. Results for Clusters 0 and 3 indicated that

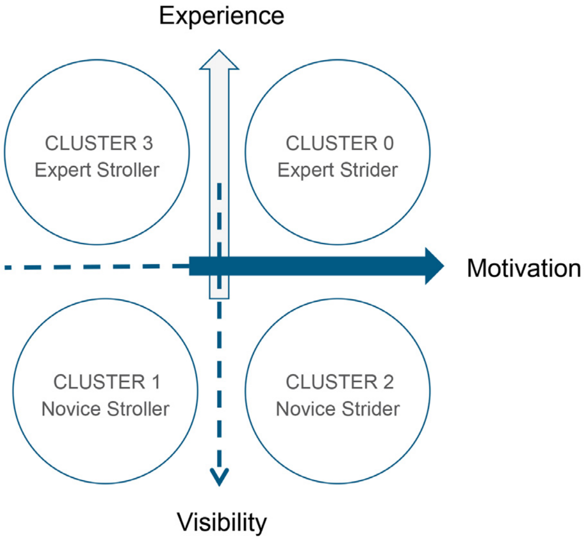

Fig. 12. Clustering results against key behaviours identified.

their users prefer moving in locations with reduced visibility, approximately 65% less than those in Clusters 1 and 2. In particular, spatial visibility for Clusters 1 and 2 has an observed mean value of up to 38K, while Clusters 0 and 3 revealed mean values of 13K.

The analysis on the number of unique devices revealed that the unique users recorded for Clusters 1 and 2 are higher by 5% in relation to the rest of the Clusters. It is hypothesised that the reason why a significant percentage of unique recorded devices falls within the rest of the clusters is because the point-by-point analysis was employed, indicating that behaviours change as someone moves within the space. Distractions and content divert pedestrians' plans to minimise distance and effort and vice versa (Al-Widyan et al., 2017). The exploratory journey may turn into a goal-oriented one when awakening of interest occurs, aiming to reach the specific destination-area of interest. In a similar way, the walking behaviour may be altered to an exploratory journey, evoked by the emotional qualities of an environment. Further to this, once the individual achieves a visual connection with their final destination, then the level of area familiarity can change from no knowledge to increased knowledge, as it enhances a faster acquisition of the cognitive map of the destination (legibility) (Zacharias, 2001). The results also revealed that the number of unique devices increase during weekend period, by 5%, while they show a 10% decrease on Monday. These results indicate that the area acts as an attractor for unique users during weekends.

Employment of BDAs and more specifically, unsupervised ML algorithms in the context of walking activities in urban space, present several challenges. These challenges include the selection of suitable datasets and variables, and to overcome these, domain knowledge is required. Pedestrian activity at a specific place and time is influenced by numerous factors related to the urban built environment, such as network connectivity, spatial visibility, and quality of view, introducing challenges in the types of data needed. Due to the high expense associated with primary data collection, data sources must be chosen carefully, to ensure the study question will be addressed.

Conclusion

The aim of this study is to present a novel means to assess pedestrian routing in urban environments. By investigating walking patterns in conjunction with spatial attributes, such as spatial visibility, insights are revealed in relation to individual preferences and behaviours of end-users and utilisation of the urban space. The academic contribution of this study centres in: (a) enhancing the body of knowledge by developing a conceptual model to assess and classify pedestrian movement behaviours, utilising machine learning algorithms and location data in conjunction with spatial attributes, and (b) extending previous research by revealing spatial visibility as a driver for pedestrian movement in urban environments. The importance of the findings lies in the perspective of revealing novel insights concerning individual preferences and behaviours of end-users and the utilisation of urban spaces. The present study adds to the wayfinding and transportation existing literature by looking upon large-scale contexts as an addition to traditional methods. The results indicate that the conceptual model developed provides a consistent path for identifying pedestrian categories across a range of different periods and it can be directly applied to urban design approaches.

By analysing the results occurred via the employment of the conceptual model (ML algorithms and location data derived from Wi-Fi tracking techniques), the authors have drawn the following conclusions/interpretations.

The Elbow Method, Silhouette Analysis and Calinski-Harabasz coefficient methods were utilised to assess the number of distinct clusters within the tested data. All the methods were tested in different resolutions (daily, period, and hourly), and returned same results, indicating the consistency in the types of pedestrians that the study area attracts. Considering the amount of people attracted in the study area daily, these results indicate that the urban design qualities and activities of an area may influence pedestrian walking patterns in a way that individual characteristics can be reduced into key common movement characteristics.

The analysis revealed four clusters for each day of the overall dataset representing the key behavioural patterns of pedestrians. The clustering results were driven by the spatial visibility parameter, as revealed by the feature importance extraction analysis, indicating the importance of that spatial attribute to pedestrian movement.

The individual characteristics of the clusters revealed, indicated that two of the four clusters spent time in areas with increased spatial visibility. Subsequently, these clusters had the higher number of unique users, indicating that knowledge of the area is an important influencing parameter of movement.

Two of the clusters recorded higher movement speed, indicating that pedestrians were motivated towards a destination. However, motivation of pedestrians did not indicate a close relation to spatial visibility and requires further investigation.

Implications/limitations of this study include the fact that due to technical issues data collection was not continuous with missing entries for some of the days. Collection of pedestrian movement in this study was restricted to a specific number of days, spanning from August to October, introducing further limitations relative to neglected spatio-temporal aspects of human behaviour. Additionally, the limited positioning accuracy of the data can introduce significant error in a streetscape when assessing walking behaviours in the context of the micro level design, highlighting the limitations of employing only one approach of data collection for such a complex study. Furthermore, the employment of unsupervised ML algorithms in the context of walking activities in urban space, presents limitations, such as model overfitting, where a model represents noise or random errors, rather than revealing actual patterns inherent in data. ML algorithms provide improved performance with more data, however, increasing model complexity can result in deterioration in performance.

Future research may focus on the analysis of additional datasets or the collation of further real-time data that could enable the investigation of variations from exogenous parameters, such as weather conditions or one-time incidents, leading to increased robustness and more accurate predictions. Moreover, exploration of data collection techniques is needed to compare such methods related to pedestrian movement tracking, such as GPS devices and video recordings. Such approaches can achieve their full potential if they are integrated in the urban design process, requiring major changes in how design is implemented. Hence fundamental changes are required in educational and behavioural contexts. Furthermore, it would be beneficial to investigate additional urban and suburban areas, such as high streets or areas near schools, comparing identified behaviours in busy major shopping streets to high streets found in lesser active urban settings. Finally, a better understanding of the key characteristics of the identified behaviours (clusters) and the spatial attributes that encourage or discourage their presence is needed. A comparison of speed profiles, occupancy patterns and weather information can provide beneficial insights. Such conceptual models could provide improved understanding of post-COVID re-evaluation of urban spaces, especially in similar contexts where retail activities are in decline. Most likely, one of the key priorities for urban city centres, as these recover, would be their economic development. Nevertheless, as urban geometry of cities vary significantly, one common approach is not easy to be adopted. Therefore, context-specific approaches, such as the conceptual model proposed in this study, are required to enable effective planning and response.

Declaration of Competing Interest

The author(s) declare no conflict of interest with respect to the research, authorship and/or publication of this article.

Author contributions

The final manuscript has been approved by all three authors. Avgousta Stanitsa conceived of the presented idea and compiled the manuscript. Stephen H. Hallett and Simon Jude supervised this work, verified the methods used, and provided substantial inputs to the text. All three authors discussed the results and contributed to the final manuscript.

Acknowledgments

This work was supported by the Natural Environment Research Council (grant number NE/R011265/1). The authors are grateful to the Natural Environment Research Council (NERC) and Economic and Social Research Council (ESRC). We are also grateful to Atkins for support. The authors wish to express gratitude to The Crown Estate for provision of Wi-Fi data for analysis and testing.

Ethical approval

An ethical approval was obtained for the collection and use of the Wi-Fi data via the Cranfield University Research Ethics System (CURES). Reference: CURES/10888/2020

Data statement

Due to the sensitive nature of the data collected and used in this study, certain data remain confidential and are not shared. Other data are available open-source as indicated.

References

Abu-Bakar, H., Williams, L., Hallett, S., 2021. Quantifying the impact of the COVID-19 lockdown on household water consumption patterns in England. npj Clean Water 4 (13).

Accuware Inc, 2017. Accuware. Accuware Inc. [Online]. Available https://accuware-inc.com/[Accessed 19 08 2021]

Ahrné, K., Bengtsson, J., Elmqvist, T., 2009. Bumble bees (Bombus spp) along a gradient of increasing urbanization. PLoS One 4 (5).

Al-Widyan, F., Al-Ani, A., Kirchner, N., Zeibots, M., 2017. An effort-based evaluation of pedestrian route choice. Sci. Res. Essays 12 (4), 42–50.

Ali, Y., Zheng, Z., Haque, M., 2021. Modelling lane-changing execution behaviour in a connected environment: a grouped random parameters with heterogeneity-in-means approach. Commun. Transport. Res. 1 (100009).

Amin and, A., et al., 2016. Comparing oversampling techniques to handle the class imbalance problem: a customer churn prediction case study. IEEE Access 4, 7940–7957.

Angelelli, F., Morrow, J., Greenwood, C. The potential application of Wi-Fi data in the development of agent based pedestrian models., in *The European Transport Conference*, Dublin, Ireland, 2018.

Aschwanden, G., Wijnands, J., Thompson, J., Nice, K., Zhao, H., Stevenson, M., 2019. Learning to walk: modeling transportation mode choice distribution through neural networks. Environment and Planning B: Urban Analytics and City Science.

Bennett, D., 2001. How can I deal with missing data in my study? Aust N Z J Public Health 25 (5), 464–469.

Bitgood, S., 2010. An analysis of visitor circulation: movement patterns and the general value principle. Curator 49 (4), 463–475.

Bohannon, R., 1997. Comfortable and maximum walking speed of adults aged 20-79 years: reference values and determinants. Age Ageing 26 (1), 15–19.

Boumezoued, S., Bada, Y., Bougdah, H., 2020. Pedestrian itinerary choice: between multi-sensory, affective and syntactic aspects of the street pattern in the historic quarter of Bejaia, Algeria. Int. Rev. Spatial Plann. Sustain. Develop. 8 (4), 91–108.

Brown, G., Schebella, M., Weber, D., 2014. Using participatory GIS to measure physical activity and urban park benefits. Landsc. Urban Plan. 121, 34–44.

Calinski, T., Harabasz, J., 1974. A dendrite method for cluster analysis. Commun. Statistics 3 (1), 1–27.

Capitanio, M., 2019. Attractive streetscape making pedestrians walk longer routes: The case of Kunitachi in Tokyo. J. Architect. Urbanism 43 (2), 131–137.

Carmona, M., 2015. London's local high streets: The problems, potential and complexities of mixed street corridors. Prog. Plann. 100, 1–84.

Celebi, M., Kingravi, H., Vela, P., 2013. A comparative study of efficient initialization methods for the k-means clustering algorithm. Expert Syst. Appl. 40 (1), 200–210.

Chang, H.-H., Chen, S., 2009. Consumer perception of interface quality, security, and loyalty in electronic commerce. Inf. Manag. 46 (7), 411–417.

Chang and, X., et al., 2020. Understanding user's travel behavior and city region functions from station-free shared bike usage data. Transport. Res. Part F 72, 81–95.

E. Choi, "Walkability as an Urban Design Problem: Understanding the activity of walking in the urban environment (Licentiate dissertation)," 2012.

Clifton, K., Smith, A., Rodriguez, D., 2007. The development and testing of an audit for the pedestrian environment. Landsc. Urban Plan. 80 (1-2), 95–110.

Colone, F., Woodbridge, K., Guo, H., Mason, D., Baker, C., 2011. Ambiguity function analysis of wireless LAN transmissions for passive radar. IEEE Trans. Aerosp. Electron. Syst. 47 (1), 240–264.

Cornell, E., Sorenson, A., Mio, T., 2003. Human sense of direction and wayfinding. Ann. Ass. Am. Geogr. 93 (2), 399–425.

Cranfield University, 16 August 2017. Supercomputer Powers up at Cranfield University. Cranfield University. [Online]. Available https://www.cranfield.ac.uk/press/news-2017/supercomputer-powers-up-at-cranfield-university[Accessed 04 06 2021]

Díaz-Álvarez, A., Clavijo, M., Jiménez, F., Talavera, E., Serradilla, F., 2018. Modelling the human lane-change execution behaviour through multilayer perceptrons and convolutional neural networks. Transport. Res. Part F 56, 134–148.

Davies, J., 2007. Yellow Book: A Prototype Wayfinding System for London. Transport for London by Applied Information Group, London.

Dridi, M., 2015. Simulation of high-density pedestrian flow: a microscopic model. Open J. Modell. Simul. 3, 81–95.

Duives, D., van Oijen, T., Hoogendoorn, S., 2020. Enhancing crowd monitoring system functionality through data fusion: Estimating flow rate from wi-fi traces and automated counting system data. Sensors (Switzerland) 20 (21), 1–25.

EUROPEAN PARLIAMENT AND OF THE COUNCIL, 2016. General data protection regulation (GDPR). Official J. Eur. Union, Brussels.

Feng, Y., Duives, D., Daamen, W., Hoogendoorn, S., 2021. Data collection methods for studying pedestrian behaviour: a systematic review. Build. Environ. 187.

Forgy, E., 1965. Cluster analysis of multivariate data: efficiency vs interpretability of classifications. Biometrics 21, 768–780.

Frank, L., 2000. Land use and transportation interaction: implications on public health and quality of life. J. Plann. Educ. Res. 20 (1), 6–22.

French, S.P., Barchers, C., Zhang, W., 2015. How should urban planners be trained to handle big data? In: Proc. NSF Workshop on Big Data and Urban Informatics. Chicago.

P. Fuxjaeger and S. Ruehrup, "Towards privacy-preserving wi-fi monitoring for road traffic analysis," 22 February 2018. [Online]. Available: https://www.researchgate.net/publication/305877717.

Gärling, T., Gärling, E., 1988. Distance minimization in downtown pedestrian shopping. Environ. Plann. A 20 (4), 547–554.

Gehl, J., Gemzoe, L, 1996. Public Spaces. Public Life., Copenhagen, Denmark: The Danish Architectural Press and Royal Danish Academy of Fine Arts. School of Architectural Publishers.

Gehl, J., Svarre, B., 2013. How to Study Public Life, 2nd Ed. Island Press, Washington, DC.

Gehl, J., 2011. Life Between Buildings: Using Public Space. The Danish Architectural Press.

Gibson, E., 1988. Exploratory behavior in the development of perceiving, acting, and the acquiring of knowledge. Annu. Rev. Psychol. 39 (1), 1–42.

Hair, J.F., Black, W., Babin, B., Anderson, R., 2014. Multivariate data analysis. Pearson Education Limited, Harlow.

Han, J., Kamber, M., Pei, J., 2012. Data Mining: Concepts and Techniques. A volume in The Morgan Kaufmann Series in Data Management Systems. Elsevier.

Harari, G., Lane, N., Wang, R., Crosier, B., Campbell, A., Gosling, S., 2016. Using smartphones to collect behavioral data in psychological science: opportunities, practical considerations, and challenges. Perspect. Psychol. Sci. 11 (6), 838–854.

Hillier, B., Hanson, J., 1984. The Social Logic of Space. Cambridge University Press, Cambridge, UK.

Hillier, B., Perm, A., Hanson, J., Grajewski, T., Xu, J., 1993. Natural movement: or, configuration and attraction in urban pedestrian movement. Environ. Planning B 20, 29–66.

Huang, Z., 1998. Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Min. Knowl. Discov. 24, 283–304.

Istrate, A.-L., Bosák, V., Nováček, A., Slach, O., 2020. How attractive for walking are the main streets of a shrinking city? Sustainability 12 (15), 6060.

Järv, O., Ahas, R., Saluveer, E., Derudder, B., Witlox, F., 2012. Mobile phones in a traffic flow: A geographical perspective to evening rush hour traffic analysis using call detail records. PLoS One 7 (11).

J, X., H, J., 2011. K-Medoids clustering. In: Sammut, C., WebbG, I (Eds.), Encyclopedia of Machine Learning. Springer, Boston, MA.

Joliffe, I., Morgan, B., 1992. Principal component analysis and exploratory factor analysis. Stat. Methods Med. Res. 1 (1), 69–95.

Karbovskii, V., Severiukhina, O., Derevitskii, I., Voloshin, D., Presbitero, A., Lees, M., 2019. The impact of different obstacles on crowd dynamics. J. Computat. Sci. 36 (100893).

Ki, D., Lee, S., 2021. Analyzing the effects of green view index of neighborhood streets on walking time using Google street view and deep learning. Landsc. Urban Plan. 205.

Koh, Z., Zhou, Y., Lau, B., Yuen, C., Tuncer, B., Chong, K., 2020. Multiple-perspective clustering of passive Wi-Fi sensing trajectory data. IEEE Trans. Big Data 1 1 -.

Kontokosta, C., Hong, B., Johnson, N., Starobin, D., 2018. Using machine learning and small area estimation to predict building-level municipal solid waste generation in cities. Comput., Environ. Urban Syst. 70, 151–162.

Krizek, K., Forysth, A., Slotterback, C., 2009. Is there a role for evidence-based practice in urban planning and policy? Planning Theory Practice 10, 459–478.

Lee, K-S., 2011. Interrogating 'digital Korea': mobile phone tracking and the spatial expansion of labour control. Media Int. Austr. 141 (1), 107–117.

Lee, J., 2020. Exploring walking behavior in the streets of New York city using hourly pedestrian count data. Sustainability 12 (19), 7863.

Li, X., Yu, L., Hang, L., Tang, X., 2017. The parallel implementation and application of an improved k-means algorithm. J. Univ. Electron. Sci. Technol. China 46, 61–68.

Li, X., Ye, R., Fang, Z., Xu, Y., Cong, B., Han, X., 2020. Uni- and bidirectional pedestrian flows through zigzag corridor in a tourism area: a field study. Adaptive Behav. 1–16.

Loukaitou-Sideris, A., 2020. Special issue on walking. Transport Reviews 4 (2), 131–134.

Lynch, K., 1960. The Image of the City. MIT Press, Cambridge ISBN-13: 9780262620017, ISBN 0262620014.

Ma, X., Wu, Y.-J., Wang, Y., Chen, F., Liu, J., 2013. Mining smart card data for transit riders' travel patterns. Transport. Res. Part C 36, 1–12.

Mansouri, M., Ujang, N., 2017. Space syntax analysis of tourists' movement patterns in the historical district of Kuala Lumpur, Malaysia. J. Urbanism 10 (2), 163–180.

Martín, J., Khatib, E., Lázaro, P., Barco, R., 2019. Traffic monitoring via mobile device location. Sensors 19 (4505).

Mauri, C., 2003. Card loyalty. A new emerging issue in grocery retailing. J. Retail. Consum. Services 10 (1), 13–25.

Medeiros, D., Neto, H.Cunha, Lopez, M., 2020. A survey on data analysis on large-Scale wireless networks: online stream processing, trends, and challenges. J. Internet Serv. Appl. 11 (6).

Mehta, V., 2009. Look closely and you will see, listen carefully and you will hear: Urban design and social interaction on streets. J. Urban Design 14 (1), 29–64.

Mendiola, L., González, P., 2021. Urban development and sustainable mobility: a spatial analysis in the Buenos Aires metropolitan area. Land 10 (2), 157.

Mercieca, J., Kaparias, I., Bell, M., Finch, E., 2011. Integrated street design in high-volume junctions: the case study of London's Oxford Circus. 1st International Conference on Access Management. Greece, Athens.

Milligan, G., Cooper, M., 1985. An examination of procedures for determining the number of clusters in a data set. Psychometrika 50 (2), 159–179.

S. Moore, "Opportunities for conversational AI in government," 2017.

Moreira, F., Ferreira, M., 2017. Teaching and learning requirement engineering based on mobile devices and cloud: a case study. In: Blended Learning: Concepts,Methodologies, Tools, and Applications, pp. 1190–1217.

Mouratidis, K., 2021. Urban planning and quality of life: A review of pathways linking the built environment to subjective well-being. Cities 115 (103229).

T. Namratha Reddy and K. P. Supreethi, "Optimization of K-means algorithm: ant colony optimization," in *2017 International Conference on Computing Methodologies and Communication (ICCMC)*, Erode, India, 2017.

Narayanan, B., Djaneye-Boundjou, O., Kebede, T., 2016. erformance analysis of machine learning and pattern recognition algorithms for Malware classification. In: Proceedings of the 2016 IEEE National Aerospace and Electronics Conference (NAECON) and Ohio Innovation Summit (OIS). Dayton, OH, USA.

Narayanan, B., Hardie, R., Kebede, T., Sprague, M., 2019. Optimized feature selection-based clustering approach for computer-aided detection of lung nodules in different modalities. Pattern Anal. Appl. 22, 559–571.

O'Sullivan, D., Morrison, A., Shearer, J., 2000. Using desktop GIS for the investigation of accessibility by public transport: an isochrone approach. Int. J. Geograph. Inf. Sci. 14, 85–104.

Özer, Ö, Kubat, A., 2015. Measuring walkability in Istanbul Galata Region. ITU A|Z 12 (1), 15–29.

Peftitsi, S., Jenelius, E., Cats, O., 2020. Determinants of passengers' metro car choice revealed through automated data sources: a Stockholm case study. Transportmetrica A 16 (3), 529–549.

Peng, C., Harwell, M., Liou, S., Ehman, L., 2006. Advances in missing data methods and implications for educational. In: Sawilowsky, S. (Ed.), Real data analysis. Information Age Pub, North Carolina, pp. 31–78.

Phillips, J., Walford, N., Hockey, A., Foreman, N., Lewis, M., 2013. Older people and outdoor environments: pedestrian anxieties and barriers in the use of familiar and unfamiliar spaces. Geoforum 47, 113–124.

Pollard, J.A., Spencer, T., Jude, S., 2018. Big Data Approaches for coastal flood risk assessment and emergency response. WIREs Clim. Change.

Reddy, G.T., 2020. Analysis of dimensionality reduction techniques on big data. IEEE Access 8, 54776–54788.

Resch, B., Puetz, I., Bluemke, M., Kyriakou, K., Miksch, J., 2020. An interdisciplinary mixed-methods approach to analyzing urban spaces: the case of urban walkability and bikeability. Int. J. Environ. Res. Public Health 17 (19).

Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65.

Scikit-learn 0.19.1 documentation, 2018. Sklearn.preprocessing.MinMaxScaler. Scikit-learn. [Online]. Available http://scikitlearn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html[Accessed 2018]

Shi, Q., Abdel-Aty, M., 2015. Big Data applications in real-time traffic operation and safety monitoring and improvement on urban expressways. Transport. Res. Part C 58, 380–394.

The Crown Estate, 2021. The Crown Estate. The Crown Estate. [Online]. Available https://www.thecrownestate.co.uk/[Accessed 19 8]

Thorndike, R.L., 1953. Who belongs in the family? Psychometrika 18, 267–276.

Till, J., 2007. Architecture Depends. The MIT Press, Cambridge (MA & UK).

Tudor-Locke, C., Bittman, M., Merom, D., 2005. Patterns of walking for transport and exercise: a novel application of time use data. Int. J. Behav. Nutr. Phys. Act 2 (5).

Turner, A., Doxa, M., O'Sullivan, D., Penn, A., 2001. From isovists to visibility graphs: a methodology for the analysis of arcthiectural space. Environ. Plann. B 28, 103–121.

Van Dijk, J., 2018. Identifying activity-travel points from GPS-data with multiple moving windows. Comput., Environ. Urban Syst. 70.

T. Varoudis, "depthmapX - multi-platform spatial network analysis software," [Online]. Available: https://varoudis.github.io/depthmapX/. [Accessed 19 08 2021].

Vinutha, H.P., Poornima, B., Sagar, B.M., 2018. Detection of outliers using interquartile range technique from intrusion dataset. In: Satapathy, S., Tavares, J., Bhateja, V., Mohanty, J. (Eds.), Information and Decision Sciences. Advances in Intelligent Systems and Computing. Springer, Singapore, pp. 511–518 vol. 701.

Wagstaff, K., Cardie, C., Rogers, S., Schrödl, S., 2001. Constrained k-means clustering with background knowledge. In: Proceedings of the Eighteenth International Conference on Machine Learning. Williamstown, MA, USA.

Wang, S.-M., Huang, C.-J., 2019. Using space syntax and information visualization for spatial behavior analysis and simulation. Int. J. Adv. Comput. Sci. Appl. 10 (4), 510–521.

Wang, Q., Wang, C., Feng, Z., Ye, J., 2012. Review of K-means clustering algorithm. Electron. Des. Eng. 20, 21–24.

Wang, J., Gao, Q., Pan, M., Fang, Y., 2018. Device-free wireless sensing: challenges, opportunities, and applications. IEEE Network 32 (2).

Weisman, J., 1981. Evaluating architectural legibility: way-finding in the built environment. Environ. Behav. 13 (2), 189–204.

W. Whyte, The social life of small urban spaces, project for public spaces, 1980.

Wirz, M., Franke, T., Roggen, D., Mitleton-Kelly, E., Lukowicz, P., Tröster, G., 2013. Probing crowd density through smartphones in city-scale mass gatherings. EPJ Data Science 2 (1), 1–24.

Ye, Y., Zeng, W., Shen, Q., Zhang, X., Lu, Y., 2019. The visual quality of streets: a human-centred continuous measurement based on machine learning algorithms and street view images. Environ. Plann. B 46 (8), 1439–1457.

Yuan, C., Yang, H., 2019. Research on K-value selection method of k-means clustering algorithm. J vol. 2 (no. 2), 226–235.

Yue, L., Abdel-Aty, M., Wang, Z., 2022. Effects of connected and autonomous vehicle merging behavior on mainline human-driven vehicle. J. Intell. Connect. Vehicles 5 (1), 36–45.

Zacharias, J., 2001. Pedestrian behavior and perception in urban walking environments. J. Planning Literature 16 (1), 3–18.

Zhai, D., Yu, J., Gao, F., Lei, Y., Feng, D., 2014. K-means text clustering algorithm based on centers selection according to maximum distance. Appl. Res. Comput. 31, 713–719.

Zhang, J., Leung, Y.-W., 2003. Robust clustering by pruning outliers. In: IEEE Transactions on Systems, Man, and Cybernetics – Part B, 33, pp. 983–999.

Zhang, N., Chen, F., Zhu, Y., Peng, H., Wang, J., Li, Y., 2020. A study on the calculation of platform sizes of urban rail hub stations based on passenger behavior characteristics. Math. Probl. Eng. 7, 1–14.

Zhu, S., Wang, D., Li, T., 2010. Data clustering with size constraints. Knowl.-Based Syst. 23, 883–889.