RESEARCH ARTICLE

WILEY **Transactions in GIS** *tg*

The characteristics of asymmetric pedestrian behavior: A preliminary study using passive smartphone location data

Nick Malleson[1] | Anthony Vanky[2,3] | Behrooz Hashemian[2] |
Paolo Santi[2,4] | Santosh K. Verma[5] | Theodore K. Courtney[2,6] |
Carlo Ratti[2]

[1] School of Geography, University of Leeds, Leeds, United Kingdom

[2] Senseable City Laboratory, Department of Urban Studies and Planning, MIT, Cambridge, Massachusetts

[3] Department of Urban Studies and Planning, MIT, Cambridge, Massachusetts

[4] Istituto di Informatica e Telematica del CNR, Pisa, Italy

[5] Department of Family Medicine and Community Health, University of Massachusetts Medical School, Worcester, Massachusetts

[6] Environmental and Occupational Medicine and Epidemiology Program, Department of Environmental Health, Harvard T. H. Chan School of Public Health, Boston, Massachusetts

Correspondence
Nick Malleson, School of Geography, University of Leeds, Leeds LS2 9JT, United Kingdom.
Email: n.s.malleson@leeds.ac.uk

Funding information
UK Economic and Social Research Council, Grant/Award No. ES/L009900/1

Abstract

Understanding the movements of people is essential for the design and management of urban areas. This article presents a novel approach to understanding the asymmetry in route choice (i.e., the degree to which people choose different walking routes for their outbound and return journeys). The study utilizes a large volume of traces of individual routes, captured using a smartphone application. The routes are aggregated to a regular grid, and matrix statistics are developed to estimate the aggregate degree of route asymmetry for different types of route (shortest, longest, weekday, weekend, etc.). The results suggest that people change their route approximately 15% of the time. Although this varied little when observing trips made at the weekend or on a weekday, people taking journeys that deviated substantially from the shortest possible path were 6 percentage points less likely to change their routes than those taking journeys that were closest to the shortest path (14 and 20% asymmetry, respectively). The absolute length also impacted on the asymmetry of journeys, but not as substantially. This result is important because, for the first time, it reports a correlation between deviation from shortest route and aggregate pedestrian choice.

1 | INTRODUCTION

Understanding the movements of people is of vital importance for the design and management of urban areas; both in terms of physical infrastructure planning and for the provision of public or private transportation. Traditionally, it has

been extremely difficult to source reliable, high-resolution information about how people use urban spaces. However, the emergence of "big" social data sources, coupled with the proliferation of sensing devices (Internet of Things) and new approaches to urban governance (Smart Cities), has made it possible to collect highly detailed information about the movements of individuals around cities. Not only can these data be used empirically (e.g., for identifying infrastructure configurations that encourage or discourage walking or cycling), but they can also be used to better understand individual travel choices and subsequently inform urban design and planning theory (e.g., Gehl, 2011; Sadik-Khan & Solomonow, 2017; Speck, 2013; Whyte, 2010).

This article presents a novel approach to understanding the choices made by individuals as they traverse, on foot, the Greater Boston area of the U.S. state of Massachusetts. The data used here were generated from a smartphone application that tracked the individual journeys of its registered users. The journeys have a spatio-temporal resolution that makes it possible to map and analyze the walking choices in some considerable detail. The overall research aim is to better understand the routes that people take as they walk in an urban area. In this article, more specifically, we are interested in studying whether aggregate routes between an origin and a destination differ from those taken in the opposite direction—termed *route asymmetry*—and the extent to which the observed asymmetry is related to other route properties such as overall route length or the degree of deviation from the shortest possible route. This is important because route asymmetry, while observed in other contexts and with different tools, has never been analyzed with such fine-grained, individual mobility traces as the ones used here. Furthermore, correlating asymmetry with other route factors can help to better understand the factors that influence pedestrian behavior and walkability. It is likely that this asymmetry implicitly reveals the route choice decisions of individuals based on the conditions of the built environment. Understanding the characteristics of these choices is important for the accuracy of many urban modeling and design applications. One example of this is *walkability* (a measure of how amenable an area is for pedestrians). Designing areas that encourage walking is an important urban design goal—for examples, see Lynch (1960), Jacobs (1961), Cervero and Duncan (2003), Ewing, Handy, Brownson, Clemente, and Winston (2006), and Forsyth, Hearst, Oakes, and Schmitz (2008)—but collecting evidence about the degree of success or failure of particular designs can be extremely difficult. Research efforts such as this offer an opportunity to begin to quantify the actual usage of urban infrastructure by pedestrians, and could ultimately help to inform planners in the creation of better urban spaces for pedestrians.

The research aims are threefold:

1. **quantify the degree of route asymmetry** (i.e., the extent to which aggregate routes differ depending on the direction of the trip);
2. **compare the degree of asymmetry to other route features to begin to estimate the factors that influence the choice to change route**;
3. conclude by **assessing the veracity of passively collected smartphone data to understand pedestrian route choices**.

The article is structured as follows. Section 2 contextualizes the study through a review of the relevant literature. Section 3 then provides a brief overview of the data, the cleaning methods used, the technique for assigning raw GPS points to real routes (map matching), and outlines the study area. It also (in Section 3.5) compares the trip lengths to their shortest-path equivalents in order to begin to unpack the relationship between (a)symmetry and path length. Section 4 conducts the bulk of the analysis by creating the required origin–destination matrices, computing the flow characteristics that are required, and quantifying asymmetry for different types of path. Section 5 draws conclusions and outlines immediate future work.

2 | RELATED WORK

While the characterization of pedestrian path symmetry has been understudied, there has been considerable prior interest in enumerating the common behaviors of trips generally. Prior studies have found that patterns of travel have surprising regularity; despite the diversity of travel history overall, humans follow simple reproducible patterns

(Ben-Akiva, Bergman, Daly, & Ramaswamy, 1984). The specific enumerations of these patterns have been of particular interest to transportation planners in understanding route choice behavior. The choices made that dictate a trip are a series of discrete and internal responses to social, economic, and physical stimuli from the urban environment around the individual (Zacharias, 2001), with their cumulative effect producing the variability in paths chosen.

Garbrecht (1970) proposed that ultimately, paths are primarily chosen so as to minimize travel distance. Despite the importance of visual or perceptive aspects, survey-based research has also shown that most people chose the shortest paths over perceived levels of congestion, safety, or visual attractions (Seneviratne & Morrall, 1985). Hillier (2015) goes further in stating that, from a behavioral standpoint, pedestrians do not favor length but rather ease of complexity in trying to reduce the number of directional changes on a route.

Collecting sufficiently comprehensive empirical data on human travel can be extremely difficult (Brockmann, Hufnagel, & Geisel, 2006). Within the realm of pedestrian studies, most have usually relied on self-reported surveys or trip diaries for their source of data (e.g., Barros, Martínez, & Viegas, 2015; Lee & Moudon, 2006). In addition to the limited capacity to collect spatial and temporal data with precision from a large population, the reliability of these data may lose detail and precision regarding the modes and locations of travel, as recall may be imprecise. Frequently, the reliability of survey-based data is threatened by the loss of actual paths traveled (Agramunt, Meuleners, Chow, Ng, & Morlet, 2016), a limited capability to collect precise data on travel start and end times, total trip duration, and destination location. Respondents may also omit trips because they do not consider them to be "transportation," or because they simply forgot to log them; individuals may consider some walking activity and short trips to be below the threshold of reporting. Specific routes may also be lost, as GPS loggers tend to lack long-term adherence by subjects, and recall tends to be limited.

The increased availability and pervasive use of connected devices is opening up new avenues for urban-scale study. Compared to prior travel survey data, sources from mobile phone locations (Calabrese, Diao, Di Lorenzo, Ferreira, & Ratti, 2013; Palmer et al., 2013; Phithakkitnukoon & Ratti, 2011), GPS devices, public transport smart cards (Liu, Hou, Biderman, Ratti, & Chen, 2009; Seaborn, Attanucci, & Wilson, 2009), wireless access points, smartphone applications (Glasgow et al., 2014; Toole, de Montjoye, González, & Pentland, 2015), and Internet of Things (IoT) devices provide researchers with new opportunities to examine individual mobility with their lower collection costs, larger sample sizes, higher update frequencies, and broader spatial and temporal coverage of city dwellers (Calabrese et al., 2013; Diao, Zhu, Ferreira, & Ratti, 2016; González, Hidalgo, & Barabási, 2008; Ratti, Pulselli, Williams, & Frenchman, 2006; Reades, Calabrese, Sevtsuk, & Ratti, 2007). This offers opportunities to better understand the generalizable patterns of human behavior—and human mobility in this study in particular—that may ultimately improve the efficacy of infrastructure and the responsiveness of urban policy. The use of mobile phones, in particular, has the added benefit of overcoming the challenges of previous pedestrian studies, where more limited use technologies such as GPS loggers and travel surveys face scalar and adherence problems (Wolf, Guensler, & Bachman, 2001).

3 | DATA AND STUDY AREA

3.1 | Study area

This study utilizes data from the area of Greater Boston, MA, as illustrated in Figure 1. The Greater Boston area houses approximately 4.7 million people and is home to the city of Boston—the capital of the state of Massachusetts—as well as Cambridge, a large adjacent city that houses Harvard University and the Massachusetts Institute of Technology (MIT). The most substantial factors that influence flows around the area include the numerous tourist attractions in Boston (it is one of America's oldest cities and played a prominent role in the American Revolution), the large universities in Cambridge, and the employment and leisure opportunities afforded by central Boston.

Figure 1 also illustrates the bounding box for the analysis area. This area includes Boston and Cambridge, as well as other adjacent cities, and extends further out into the suburbs. Although the full dataset of mobility traces (discussed below) spans a somewhat larger area, the data volume decreases rapidly with distance from the center of the city of Boston, so we focus predominantly on the broad area that has the largest number of trajectories.
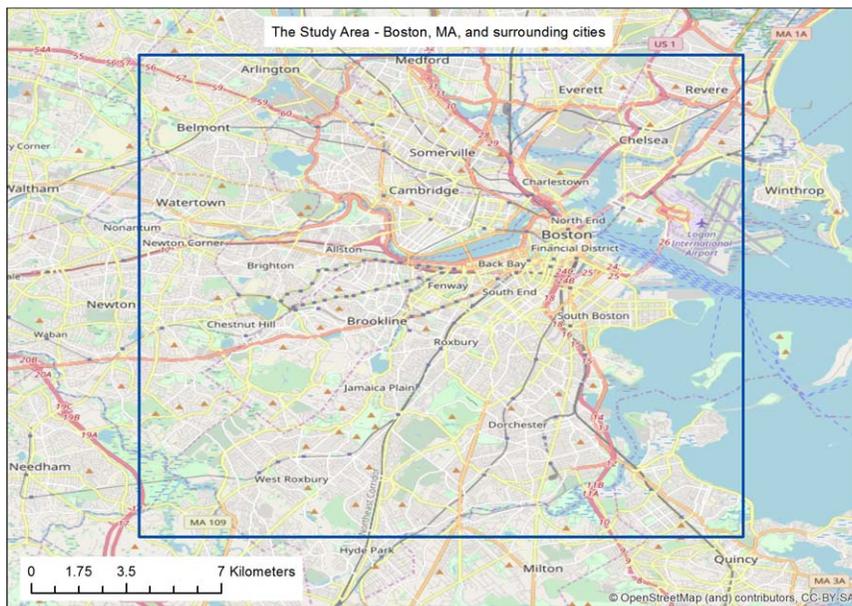
FIGURE 1 The study area: Boston, MA and surrounding cities

3.2 | Passively collected smartphone data

Trip activities were collected from a smartphone-based, activity-oriented mobile application.[1] The application was explicitly marketed as allowing users to track their daily routines, and users were required to actively install the application (it was not installed as a side-effect of some other activity). The application runs in the background of the phone's operation, automatically keeping track of the user's movements using the device's motion co-processor to record the time and movements of the phone. Geographic information was assigned to those activities through the use of a device's geolocation services, including assisted GPS, tower-based positioning, and satellite-based GPS.

A trip is defined as when a user departs a geo-fenced area of their current location (the start) until they remain in another location for a duration of time, as determined by the application's proprietary stay-detection algorithm. As this process occurred in the background, information on the users' movements was passively recorded and allows for a more complete documentation of users' trips than if they had to input the trips manually. The data covered a total time period from May 15, 2014 through May, 1 2015. In sum, the application recorded 263,670 trips from 6,424 unique users in the case study area.

Interestingly, there is some temporal regularity to the trip data. As illustrated in Figure 2, trip frequencies peak in the morning, at lunch, and again in the afternoon. This suggests that the data might adequately represent typical "9-to-
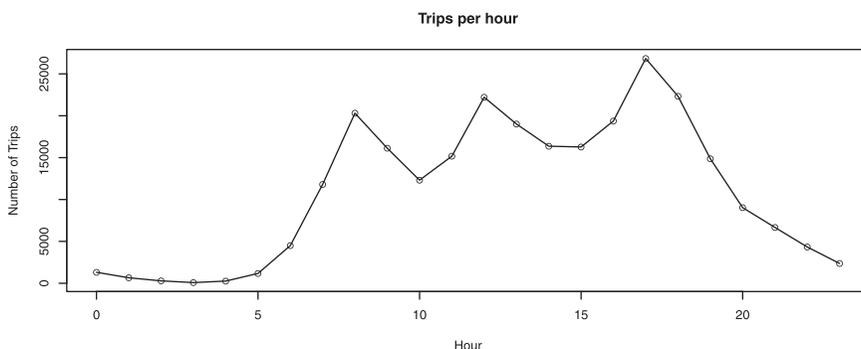

FIGURE 2 The starting times of all trips by hour of day

5" working patterns. This is important, because the aim here is to better understand "typical" routes; therefore, the similarity with a "typical" working day schedule is encouraging. It would of course be possible to aggregate the data by hour or by day and begin to analyze the routes taken at different times. This is broadly beyond the scope of this article, although (as Section 4.6 will discuss), some preliminary analysis of the trips made during weekdays versus weekends will be performed.

3.3 | Initial data cleaning

The aim of this work is to ascertain the degree of symmetry in route choices. Therefore it is not only necessary to clean and filter *unusual* trips that might be indicative of noise in the data (i.e., those that are extremely short or unrealistically long), but it is also necessary to remove those that were not made for the purpose of traveling to a destination. The software application that was used to collect the data was frequently used to track circular journeys that were presumably made for leisure or fitness purposes. As these journeys were made for their own sake, not necessarily to get to a particular destination, they cannot be used to assess the degree of route (a)symmetry. Also, there were a number of trips with unrealistically few steps given the total journey time that were also assumed to be noise. Therefore, traces were removed that met the following conditions:

- the distance between the origin and the destination was less than 300 m (circular trips or short trips);
- the elapsed time was less than 13 minutes (this was chosen because 13 minutes is approximately the value of the first quartile (Q1) of all trip durations);
- the trace contained fewer than 500 steps.

After cleaning, 103,835 individual traces remained in the analysis, compared to 263,670 traces initially.

3.4 | Map matching

The raw trip data are represented as a list of latitude/longitude coordinates, each with an associated timestamp. Before it is possible to begin analyzing the routes taken by individuals, it is necessary to assign the points to the road network and to generate a list of path segments that represent the trace (a "route"). The problem of matching coordinates to a map (road objects in this case) is often termed *map matching*, and has been well studied; for an overview, the interested reader can refer to Ahmed, Karagiorgou, Pfoser, and Wenk (2015). Here, the algorithm proposed by Newson and Krumm (2009), as implemented in the GraphHopper Map-Matching library (https://github.com/graphhopper/map-matching), was used. The approach utilizes a hidden Markov model (HMM) to find the road segments that are the *most likely* to characterize the route, given the locations of the individual GPS coordinates. An advantage with the Newson and Krumm (2009) algorithm is that it incorporates both time and space into its likelihood estimate. Therefore, noisy GPS points that might otherwise cause a wide route deviation are unlikely to be problematic if they are an unrealistic distance (in time and space) away from the most likely route. The GraphHopper library also includes routines to calculate the shortest path, so the shortest path is computed simultaneously (this is used in later analysis). OpenStreetMap data defined the road network used for map matching. As we assume that traces represent walking routes, the GraphHopper library constructs a graph representation of the road network which only includes paths that are suitable for pedestrians. The use of these data has the advantage that diverse pedestrian routes, such as those that pass through parks and leisure centers, should be included as possible path choices. A drawback is that the data are not of sufficient resolution to match to particular sidewalks or to specific crossings over major roads. This limits the accuracy of the resulting matched paths, but it will not cause any substantial issues in this research as the analysis is conducted at an aggregate level anyway.

Overall, 94,760 GPS traces were successfully matched to the OpenStreetMap network. Some could not be matched because the algorithm could find no route from the origin to the destination. Fortunately these unmatched

(a) Successful Map Matching
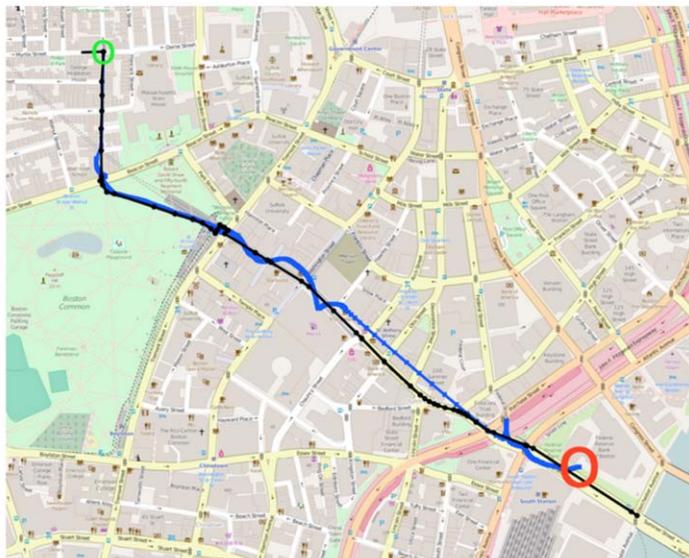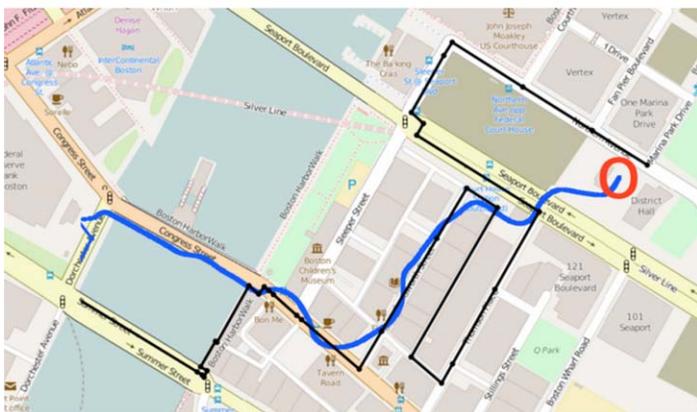

(b) Un-uccessful Map Matching

FIGURE 3 An example of successful (a) and unsuccessful (b) map matching. Raw GPS coordinates are blue, the matched path is black. In (b) it appears that the user traveled through a shopping center and car park that had no explicit paths in the OpenStreetMap data, hence the algorithm was not able to match the route effectively

routes are few (9% of all cleaned traces), and they are likely to be spurious anyway (e.g., starting or finishing over water). Unmatched paths are not included in the analysis.

By comparing the distances of the matched paths to those of the original (raw) paths, it became apparent that a small number of paths failed to adequately capture the path that is represented by the GPS coordinates. See Figure 3 for an example. It is therefore necessary to attempt to remove some of these poorly matched paths. Whilst the shortest paths and the matched paths will often be different—there is no reason to think that people will always use the shortest path, given a choice set of alternatives (Bekhor, Ben-Akiva, & Ramming, 2006; Ben-Akiva et al., 1984)—they should not be wildly different. Two error measures were used to quantify the differences in distance between the original paths and their comparative matched paths: the absolute difference, *absDiff* and the relative difference, *relDiff*:

$$absDiff = \sum_{i=0}^{n_t} |o_i - m_i| \tag{1}$$

$$relDiff = \sum_{i=0}^{n_t} \frac{|o_i - m_i|}{o_i} \tag{2}$$

where o_i is the total distance of the original (raw) trace i, m_i is the distance of the matched trace, and n_t is the total number of matched traces. To disregard the poorest matches, all of those considered as an outlier in either measure using the inter-quartile range were removed:

$$*Diff > Q_3 + 1.5 \times IQR$$

where IQR represents the inter-quartile range and Q_3 denotes the third quartile. This left 84,955 traces generated by 4,074 users.

3.5 | Comparing trip distances

Having matched the original traces to the OpenStreetMap road network, it is possible to compare the actual routes traveled (the matched paths) to the corresponding shortest paths. This comparison provides some useful validation of the map-matching results, and affords an opportunity for further cleaning. The shortest paths were calculated using Dijkstra's shortest-path algorithm, as implemented in the GraphHopper library. Figure 4 illustrates the difference in the distances between the original paths and the shortest paths. As expected, the majority of paths are longer than their "shortest" counterparts. This is consistent with previous research, which suggests that pedestrians do not necessarily take the shortest (in terms of distance) path to a destination. This difference in behavior has been explained through a variety of lenses: a lack of awareness of the most optimal route (Helbing, 1991); hedonic motivations for attractiveness (Helbing, Molnár, Farkas, & Bolay, 2001; Millonig & Schechtner, 2007); cognitive awareness and psychological anchoring to certain landmarks (Chown, Kaplan, & Kortenkamp, 1995; Golledge, Smith, Pellegrino, Doherty, & Marshall, 1985; Sadalla, Burroughs, & Staplin, 1980); environmental factors (Cools, Moons, Creemers, & Wets, 2010); or due to the guidance of navigational tools (Streeter, Vitello, & Wonsiewicz, 1985).

Figure 4 also plots the distribution of Hausdorff distances between matched paths and shortest paths. The Hausdorff distance is a measure of geometric similarity; it is the longest straight-line distance that must be traveled from one path to reach the other. The shortest paths are different from the original paths, but in most cases the differences are small. This is an interesting finding in its own right. Exploring this deviation further is beyond the scope of this article, so the following section will move on to analyze the aggregate flow characteristics in more detail and ultimately quantify the degree of asymmetry in trip choices.

4 | FLOW CHARACTERISTICS

The article will now begin to explore the characteristics of the trips in more detail. It begins by exploring the flows in terms of trip volume and direction in each cell. Subsequently, the analysis attempts to quantify the *symmetry* of flows in order to better understand the route choices that people make.

4.1 | Calculating the origin–destination matrix

In order to compare the traces, an origin–destination (O–D) matrix is created as discussed below (following Phithakkit-nukoon & Ratti, 2011). This requires that the data are first aggregated to a regular grid so that the flows between each grid cell can be calculated. Here, a regular grid is defined with cells that have a square area of 0.73 km². This cell size was chosen because it helps to compensate for uncertainty in the location estimates (particularly as the precise origins and destinations are not known) and still maintains adequate spatial accuracy. Section 4.7 will demonstrate that the analysis is not overly sensitive to this grid resolution.

Subsequently, an O–D matrix (a form of transition matrix) can be calculated. Each element $f(i, j)$ in the matrix F quantifies the flow between two cells $i, j \in \{1, 2, 3, \ldots, N\}$, where N is the total number of cells:

Relative (proportional) difference in distances travelled

Relative difference (matched trace – shortest trace)
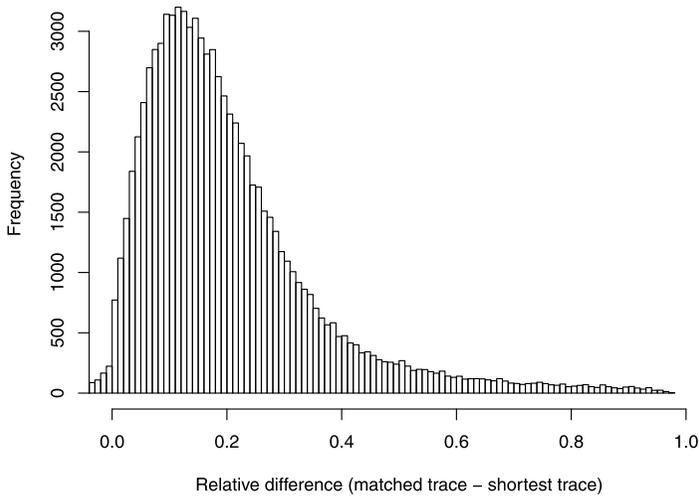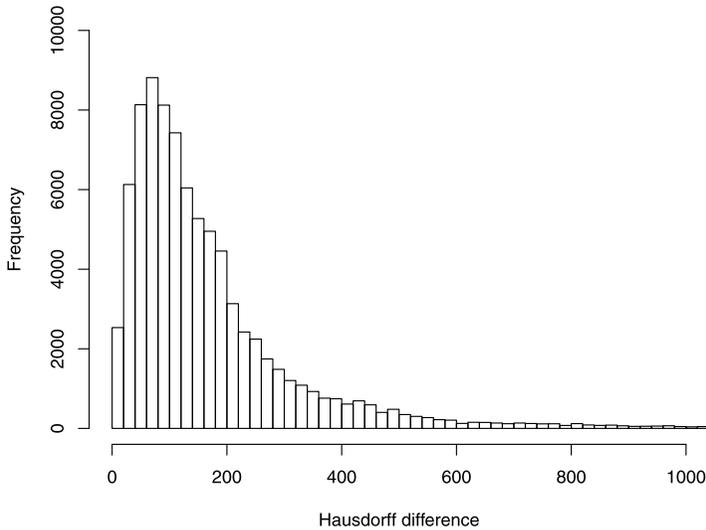
Hausdorff path difference

Hausdorff difference

FIGURE 4 Comparing the shortest paths to the original paths

$$F = \begin{bmatrix} f(1,1) & f(1,2) & \dots & f(1,N) \\ f(2,1) & f(1,2) & \dots & f(2,N) \\ \vdots & \vdots & \ddots & \vdots \\ f(N,1) & f(N,2) & \dots & f(N,N) \end{bmatrix} \tag{3}$$

It is then possible to calculate the total number of flows out of ($f_{out}(i)$) and into ($f_{in}(i)$) a cell i by summing the row and column flows, respectively:
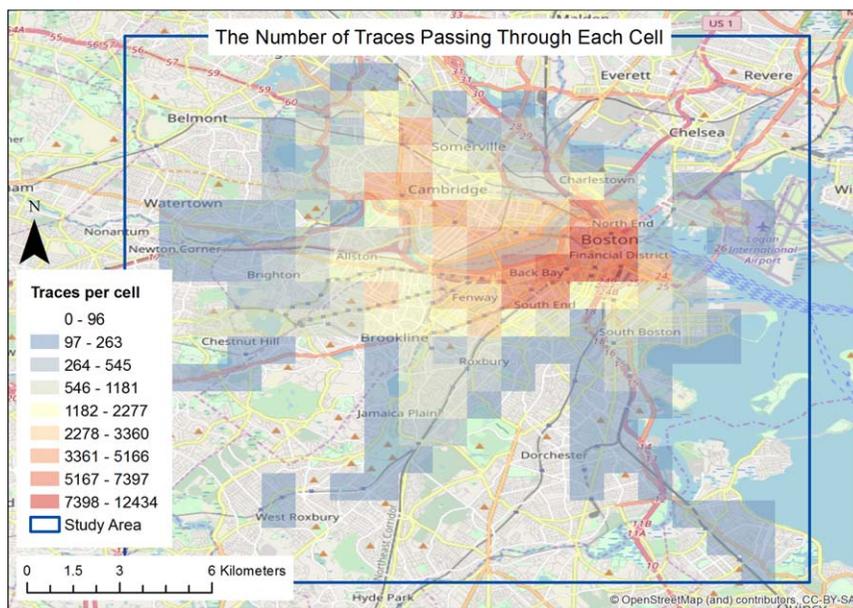
$$f_{out}(i) = \sum_{k=1,k\neq i}^{N} f(i,k) \tag{4}$$

FIGURE 5 The total flow f_{tot} passing through each cell i

$$f_{in}(i) = \sum_{k=1, k \neq i}^{N} f(k, i) \qquad (5)$$

As the original GPS points have been matched to the road network, there are no breaks in the routes. Therefore, each interaction with a cell i must be with a cell that is one of its von Neumann neighbors[2] (i.e., those four cells that share an edge with i). It is worth noting that this is an unusual form of O–D matrix. Typically, O–D matrices only measure a single interaction per journey and disregard the actual route taken. Here, however, an interaction is generated each time a trace passes through a cell. This results in multiple interactions per journey (i.e., multiple entries in the O–D matrix per journey) and *no* interactions between cells that are not adjacent. Therefore, most cells in F are 0.

4.2 | Aggregate flow characteristics

Having calculated the O–D flow matrix F, it is possible to examine the aggregate flows before moving on to an analysis of trip symmetry. Figure 5 illustrates the regular grid and the total number of traces that intersect each cell (i.e., $f_{tot}(i) = f_{in}(i) + f_{out}(i)$). The results are as expected, given the socio-spatial structure of Greater Boston. The city of Boston itself accounts for the largest number of flows, which corresponds well with the extremely large number of employment and leisure opportunities, as well as the density of the population that is present in the area. The city of Cambridge also accounts for a large number of flows, particularly around the universities. This expected pattern is encouraging, as it indicates that the traces are consistent with the expected volume of people who would be traveling in the area. Further validation could be accomplished through a comparison with other data, such as those from the U.S. Census, but we leave this for future work.

4.3 | Trip symmetry

Having provided an overview of the aggregate flow characteristics, we now explore the symmetry of the trips. A *trip*, in this context, can be defined as a journey from a starting location (i.e., "home"), through a series of intermediate cells to a destination, eventually returning to the starting location. A symmetrical trip is one where the same intermediate cells are used on the return journey, albeit in the opposite order. Figure 6 presents a time-geographic (Hägerstrand,
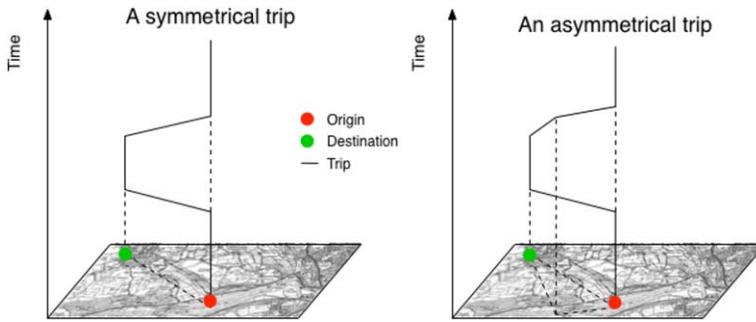
FIGURE 6 A time-geographic (Hägerstrand, 1970) representation of a symmetrical and an asymmetrical trip. Both trips involve an outbound and a return journey. In a symmetrical trip the same path is used for both trips, whereas for an asymmetrical trip different paths are used

1970) representation of a symmetrical and an asymmetrical trip. This section will begin by exploring the aggregate degree of trip asymmetry before comparing this with the lengths of trips.

The first test for symmetry can be formulated by examining the origins and destinations of flows for each cell. As all flows into and out of a cell i involve an interaction with only one of the four possible von Neumann neighbors l of i, $l(i)=\{l_1, l_2, l_3, l_4\}$, then the flows into and out of i ($f_{in}(i)$ and $f_{out}(i)$, respectively) can be reformulated as vectors in which each element represents the flow between i and l:

$$f_{in}(i)=[m(i,l_1) \quad m(i,l_2) \quad m(i,l_3) \quad m(i,l_4)] \tag{6}$$
$$f_{out}(i)=[m(l_1,i) \quad m(l_2,i) \quad m(l_3,i) \quad m(l_4,i)] \tag{7}$$

where $m(i,l_x)$ represents the flow between cell i and one of its von Neumann neighbors.

Subsequently, the relative flow $f_{rel}(i)$ can be defined as the magnitude difference (defined using Euclidean length, a scalar quantity) between the $f_{in}(i)$ and $f_{out}(i)$ vectors:

$$f_{rel}(i)=|f_{in}(i)-f_{out}(i)| \tag{8}$$

Therefore, $f_{rel}(i)$ is a single number that provides some information about the relative symmetry for an individual cell. If a trace has a symmetrical partner (i.e., a return path that takes the same route as the outbound path), then the number of traces entering and leaving the cell from a particular direction will be equal. For example, consider the case of a cell a that has one trace running through it from north to south. If the $f_{in}(a)$ and $f_{out}(a)$ vectors are specified such that the elements represent counts in the direction order north, east, south, and west, then:

$$f_{in}(a)=[1 \quad 0 \quad 0 \quad 0] \tag{9}$$
$$f_{out}(a)=[0 \quad 0 \quad 1 \quad 0] \tag{10}$$

In this case:

$$f_{rel}(a)=2 \tag{11}$$

so the flows passing through the cell are not symmetrical.

However, consider another cell (b) that has a north–south trace as well as a companion (symmetrical) trace returning from south to north. In this case the vectors become:

$$f_{in}(b)=[1 \quad 0 \quad 1 \quad 0] \tag{12}$$
$$f_{out}(b)=[1 \quad 0 \quad 1 \quad 0] \tag{13}$$

and

$$f_{rel}(b)=0 \tag{14}$$

(i.e., all traces passing through cell b have symmetrical partners).

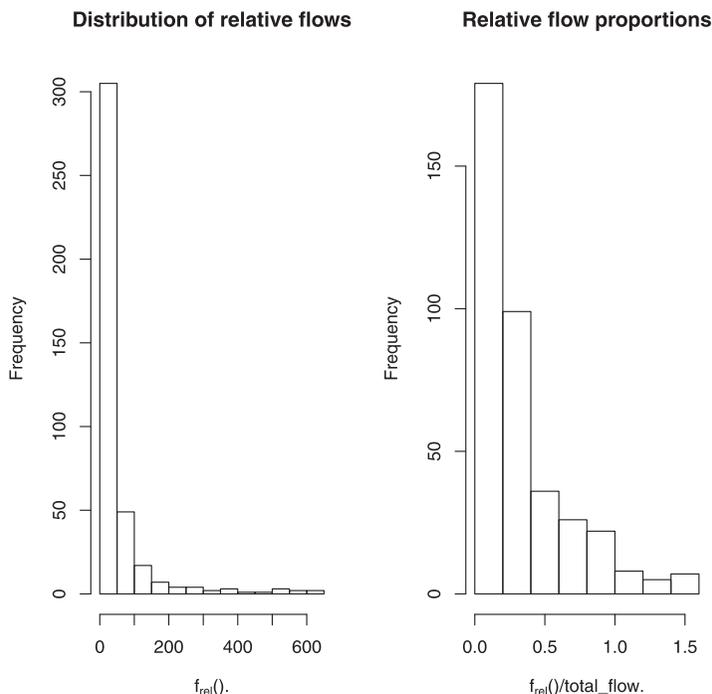**Distribution of relative flows**   **Relative flow proportions**

FIGURE 7 The relative flow in each cell ($f_{rel}(i)$). This is the difference in the directions taken by flows entering and leaving the cell. In effect, larger values mean that for traces entering the cell from a particular direction, there is no "reverse" (symmetrical) trace that enters the cell from the opposite direction

If all traces are symmetrical, then for all cells the number of traces entering the cell from each direction will be equal to the number of traces leaving from the same direction (i.e., $\forall i : f_{rel}(i)=0$). To test this, $f_{rel}(i)$ was calculated for all cells; Figure 7 illustrates the distribution of relative flows for each cell. The right-hand chart normalizes the total number of flows passing through the cell so as not to unfairly bias busy cells.

The $f_{rel}(i)$ values provide some evidence of symmetry. They suggest that most cells have a similar number of traces entering and leaving from each direction. Although Figure 7 illustrates that some cells have comparatively high $f_{rel}(i)$ values, these largely represent cells that are in the center of Boston and have an extremely large number of flows passing through them, which goes some way to explaining the differences. After normalizing by the total number of traces passing through each cell, it becomes clear that the cells with the greatest asymmetry are in the periphery of the study area and are susceptible to boundary conditions. Most cells farther inside the study area have low relative flows in comparison with the total flow count. Figure 8 illustrates this.

An advantage with the $f_{rel}(i)$ values is that because they are associated with individual cells, they can be drawn on a map directly (as per Figure 8). However, a disadvantage is that the values are difficult to interpret in an absolute sense. They make it possible to compare the differences in asymmetry across all cells, but provide no information about the proportion of flows that are (a)symmetrical. To do this, an alternative means of estimating the total aggregate asymmetry has been developed. Following Phithakkitnukoon and Ratti (2011), a *difference matrix D* is constructed:

$$D=\begin{bmatrix} d(1,1) & d(1,2) & \dots & d(1,N) \\ - & d(1,2) & \dots & d(2,N) \\ - & - & \ddots & \vdots \\ - & - & - & d(N,N) \end{bmatrix} \quad (15)$$
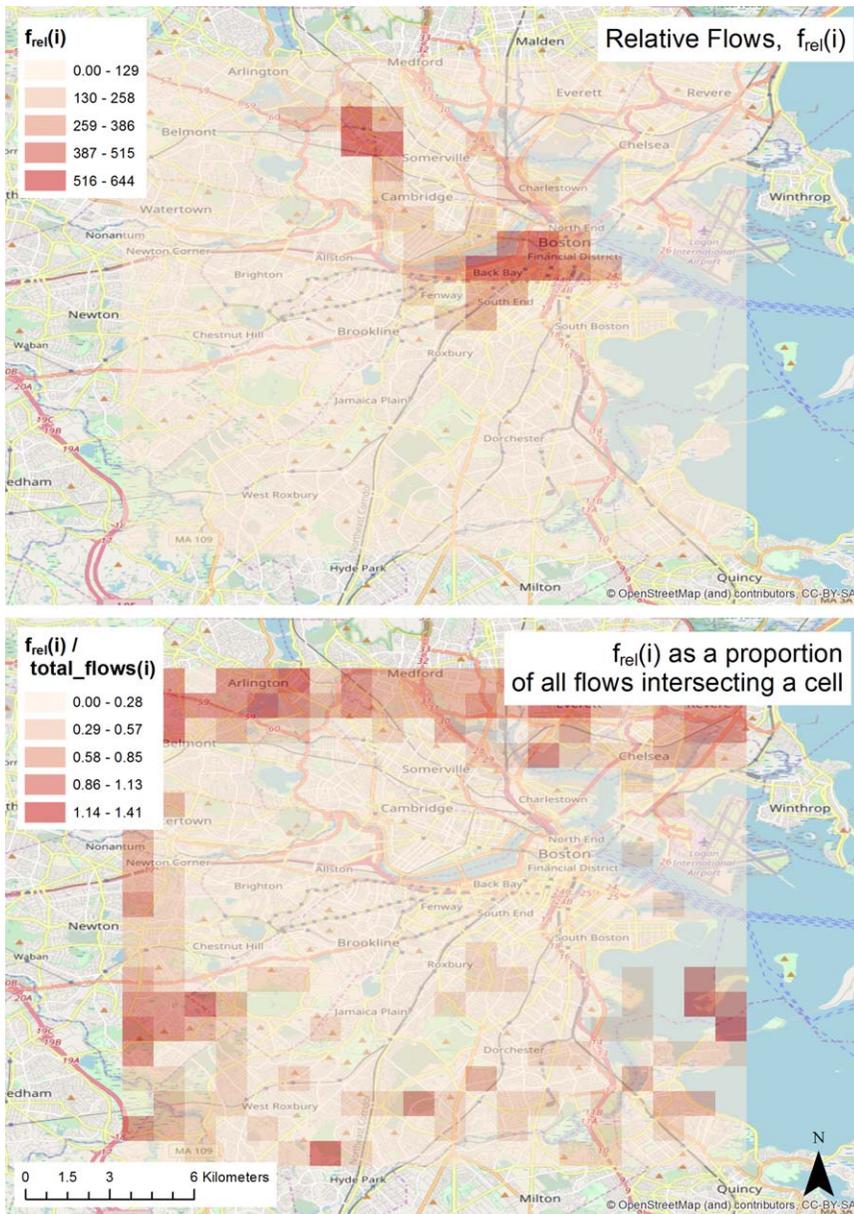
where

FIGURE 8 The relative flow ($f_{rel}(i)$) of each cell. The bottom map has been normalized by the total number of traces passing through each cell and suggests that the most asymmetric cells are susceptible to boundary influences

$$d(i,j) = |f(i,j) - f(j,i)| \qquad (16)$$

Only the upper half of the triangular matrix is required, as the bottom half is a mirror of the top half. All elements in D that are non-zero have asymmetric flows. Although there are 160,000 possible interactions (the results grid contains $20^2 = 400$ cells, so the transition matrix M contains $400^2 = 160,000$ interactions), most of these are between cells that *cannot* interact, as interactions only occur between a cell and its von Neumann neighbors. In total, there are at most four interactions per cell, so $20^2 * 4 = 1,600$ possible interactions in the transition matrix (ignoring boundary conditions) and, therefore, only 800 in D as only half of the triangular matrix is used. Of these, 611 have non-zero flow:

$$\sum_{\forall i,j: i \neq j} [d(i,j) \neq 0] = 611 \tag{17}$$

and are therefore not symmetrical. This appears to be a relatively high proportion, but does not take the size of the flows into account; cells with a large number of transactions are unlikely to be entirely symmetrical. To account for this, the *relative difference matrix* D_R is defined in the same way as D, with each element $D_R(i,j)$ defined as:

$$D_R(i,j) = \frac{d(i,j)}{\max(f(i,j), f(j,i))} \tag{18}$$

The range of $d_R(i,j)$ is $[0, 1]$, where 0 implies perfectly symmetrical flow and 1 indicates entirely asymmetrical (unidirectional) flow. These values can be interpreted as a percentage. For example, if $d_R(i,j) = 0.1$ then 10% of the flow between i and j is asymmetrical. The *overall relative asymmetry* (*Asym*; Phithakkitnukoon & Ratti, 2011) across all cells can then be calculated by summing the relative difference across all cells in the upper-triangular matrix and dividing by the sum of the maximum flows:

$$Asym = \frac{\sum_{\forall i,j: i \neq j} d(i,j)}{\sum_{\forall i,j: i \neq j} \max(f(i,j), f(j,i))} \tag{19}$$

The percentage asymmetry here is 14.78%. Therefore, asymmetric flows account for approximately 15% of all flows, or people change their routes 15% of the time. This compares to 33% as estimated by Phithakkitnukoon and Ratti (2011), which, although higher, is comparable. Explanations for these differences will be discussed in Section 5.

To summarize, we have thus far found that:

1. people tend to walk along paths that are 20% longer (on average) than the equivalent shortest path (Figure 4);

2. there is a certain degree of asymmetry in aggregate routes (people appear to change their routes approximately 15% of the time).

What remains unclear, however, is whether there is a connection between people taking paths that deviate from the shortest possible path (1) and people changing their routes (2). In other words, is the decision to take a path that deviates from the shortest route associated with the decision to take a different route to and from a destination? In the following sections, we explore this association in more detail.

4.4 | Deviation from shortest path asymmetry

In this section, we attempt to understand whether shortest path and asymmetry are related or independent choices. To answer this question, the traces were first segregated depending on the amount of deviation from their shortest-path equivalent. Those traces that are 20% longer than their shortest-path equivalent (i.e., those that deviate the most) were extracted. 20% was chosen because this represents approximately half of the traces. The traces with the greatest deviation D and the least deviation D^C can therefore be defined as:

$$D = \{t : length(t) \geq length(t_s) \times 1.2\} \tag{20}$$

$$D^C = \{t : length(t) < length(t_s) \times 1.2\} \tag{21}$$

where t is a trace, $length(t)$ is the length of the trace t, and t_s is the shortest path connecting the endpoints of t.

This results in 41,833 traces with the most deviation ($|D| = 41,833$) and 43,122 with the least ($|D^C| = 43,122$). Figure 9 illustrates the proportions of traces in D and D^C that pass through each cell in the study area. Although there are slight differences in the spatial distribution of the traces, there is no substantial visible difference in the locations of the traces in D and D^C. This implies that there is no noticeable spatial bias in the analysis relating asymmetry with deviation from the shortest path.

To properly quantify the differences in asymmetry, the percentage asymmetry (*Asym*) as defined in Equation 19 was recalculated for the subsets of traces in D and D^C. The results here are much more striking.
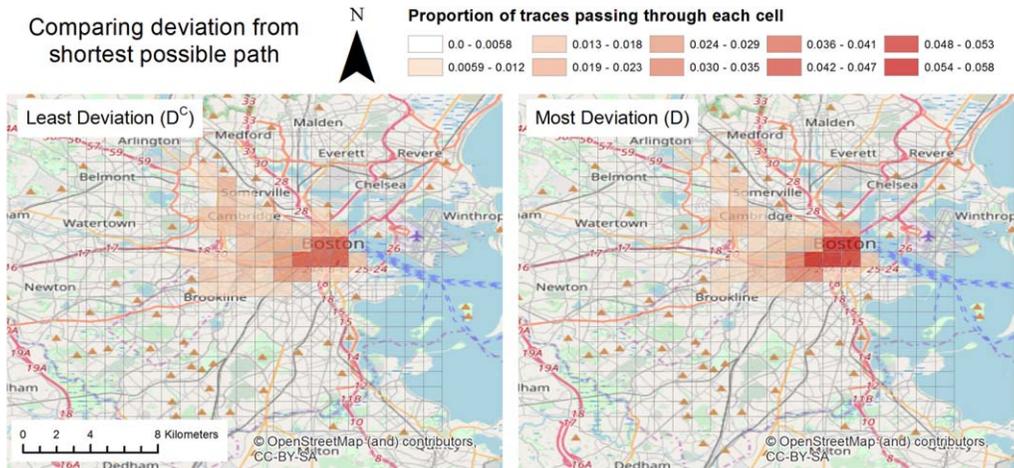
FIGURE 9 The proportions of traces with the most deviation from their shortest path (D) and least deviation (D^C) passing through each cell in the study area

- For the traces with the most deviation from their shortest-path equivalent (D), the percentage asymmetry is 13.56%. This is only marginally lower than the 14.78% asymmetry that is exhibited by *all* routes.

- For the traces with the least deviation (D^C) (i.e., those that are less than 20% longer than their shortest-path equivalents), the percentage asymmetry increases to 20.14%. This means that people who take journeys that are close to the shortest possible path are 6.6 percentage points more likely to change their route.

The above findings indicate that deviations from shortest-path routes are likely to be caused by substantial factors influencing walkability, and that those factors influence pedestrian choices independently of the direction of the route. While identifying these factors is beyond the scope of this article, the results presented herein are important since they suggest, for the first time, a statistically significant correlation between deviation from shortest route and aggregate pedestrian choices.

4.5 | Absolute path length asymmetry

Although the previous section provides some evidence that people are more likely to change their route when they travel close to the shortest possible path, it is unclear to what extent these decisions are influenced by the absolute length of the paths. For example, although an individual might know their neighborhood reasonably well, and hence have the confidence to change their path on short routes, they might always use a known route on longer trips that pass through neighborhoods they do not know so well. To better understand the relationship between absolute path length and asymmetry, the paths were separated into two equally sized subsets, L and L^C, based on the absolute length of the paths (longest and shortest, respectively).

The subsets are calculated as follows:

$$L = \{t : length(t) \geq median(T)\} \quad (22)$$
$$L^C = \{t : length(t) < median(T)\} \quad (23)$$

where t is a trace and *median(T)* is the median length of all traces. The median is used over the mean because the trip lengths are skewed. The percentage asymmetry (*Asym*), as defined in Equation 19, was then recalculated for the subsets of traces in L and L^C. Interestingly, although there is a difference in symmetry, it is less pronounced than that when the deviations from the shortest possible path are compared:

- asymmetry for absolute longest routes, 15.86%;

- asymmetry for absolute shortest routes, 18.04%.

These differences are not substantial, which is important. They show that the asymmetry associated with deviation from a shortest path is *not* overly influenced by the absolute length of the path. In other words, the decisions that cause asymmetry in trips that are a similar length to their equivalent shortest path are not related to the overall distance traveled. The implications for these findings will be discussed in Section 5.

4.6 | Day of week asymmetry

Thus far, we have found empirical evidence which suggests that people tend to walk along paths that are 20% longer (on average) than the equivalent shortest path, and that they are slightly more likely to change their route when taking shorter trips. What remains unclear, however, is whether the observed results are simply an artefact of the day that the trip is undertaken. Are trips on the weekend, for example, driven by leisure motivations rather than the need to commute, and therefore longer and less symmetrical than weekday trips?

To explore this final question, the asymmetry analysis was repeated on two subsets of trips: those that begin during a weekday, and those that begin during a weekend. However, the results did not highlight any substantial difference: the asymmetry for weekdays and weekends was 15.05 and 16.32%, respectively, compared to 14.78% for all trips. Therefore, the results appear to suggest that although deviation from shortest paths affects the symmetry of trips, there is only a small difference between trips taken on a weekday compared to the weekend.

4.7 | Sensitivity to grid resolution

It is important to note that the results are sensitive to the size of the cells chosen for the analysis. A spatial grid with smaller cells (larger resolution) will inevitably produce higher asymmetry values as it is less likely that a person will travel through the same cells on both the outbound and return trips. To test the sensitivity of the work to the underlying grid resolution, Figure 10 illustrates the different asymmetry values that arise from a number of grids of different resolution. Although asymmetry inevitably increases with resolution, the increases are not substantial. For example, if the resolution of the grid is doubled from 400 cells to 800 cells (an increase of 20 cells per row to approximately 28), the asymmetry value only increases from 15 to 17%. Furthermore, all changes in asymmetry will be in the same direction, so comparisons of asymmetry with short and long paths, or with weekday and weekend trips, still hold.

5 | DISCUSSION AND CONCLUSIONS

This article has leveraged large-volume, high-resolution route trace data in order to better understand the movements of people in urban areas. The study focuses on the area of Greater Boston, MA, but the methods are generalizable. Specifically, the article has attempted to quantify *route asymmetry*, defined as a route in which the outbound trip is different from the inbound trip. Although work on asymmetry is not new, this work is novel in that it represents the first attempt to quantify levels of asymmetry using high-resolution individual-level mobility traces. A greater understanding of the degree of asymmetry in route choice has the potential to lead to a better understanding of the factors that influence walkability and ultimately inform many urban modeling and design initiatives. To aid the discussion, Table 1 outlines the asymmetry values calculated throughout the analysis.

The article analyzed a total of 84,955 trips (after cleaning) and found that people changed their route 15% of the time. This compares to 33% as estimated by Phithakkitnukoon and Ratti (2011), although in that study the trips were estimated purely on the mobile telephone mast locations and, as such, were of substantially lower spatial accuracy. The resolution of the grids used in both studies varies as well, and as Section 4.7 illustrated, changing the resolution of the grids will influence the degree of asymmetry. The more striking finding, however, is that the degree of asymmetry appears to vary with the degree of deviation from the shortest possible path. For the subset of the traces with the greatest deviation in the dataset (those paths that are at least 20% longer than their shortest-path equivalent), people tended to change their route 14% of the time, whereas with the routes that are closer to the shortest possible path this value rose to 20%. The absolute trip length also impacted on asymmetry, but not by much (16% for the longest
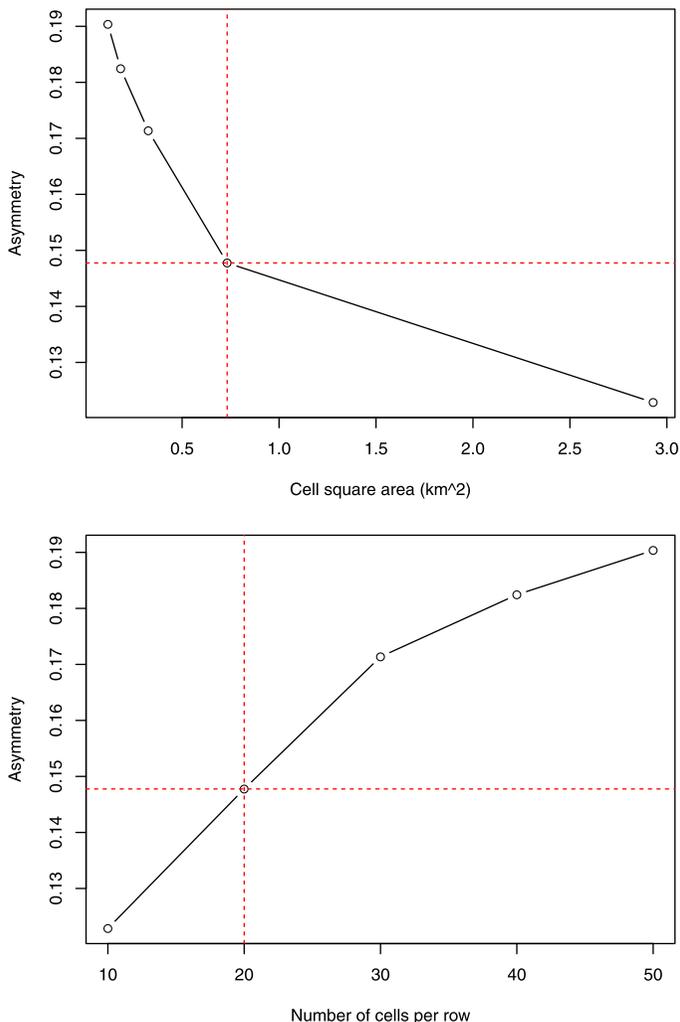
FIGURE 10 The sensitivity of the results to the resolution of the analysis grid. Red lines indicate the resolution chosen for the analysis here

paths compared to 18% for the shortest). These findings are not artefacts of the day that the trip is taken; weekday and weekend trips exhibit similar levels of asymmetry. It is possible that these differences might have been caused by personal or behavioral factors that influence route choice. For example, if someone is familiar with a route and with the neighborhoods that the route passes through, then they might be able to take slightly different routes of similar overall length without becoming lost. Conversely, if the person is *not* familiar with a route and knows only one way to and from a destination, which also happens to be non-shortest, then they might not have the required knowledge to vary their path. These are purely hypotheses, however, and it is beyond the scope of this article to dissect the complex relationship between cognition, behavior, and the physical infrastructure that might cause people taking shorter routes to vary their journeys more. This result is important nonetheless because, for the first time, it reports a correlation between deviation from shortest route and aggregate pedestrian choices. Ongoing future work will begin to qualitatively explore the specific areas that exhibit the greatest degree of asymmetry in more detail (e.g., those highlighted in Figure 8) and begin to dissect the trips themselves at an individual level.

The data used here have a number of advantages over those that have been traditionally used to elucidate information about people's journeys. The ubiquity of passively collected smartphone data, such as those used here, are particularly attractive due to their relatively low collection costs, generally larger sample sizes, regular update frequencies

TABLE 1 Asymmetry values calculated throughout the article

Section	Asymmetry calculation	Asymmetry value (%)
4.3	Overall asymmetry	14.78
4.4	Most deviation from shortest path (D)	13.56
4.4	Least deviation from shortest path (D^C)	20.14
4.5	Longest paths (L)	15.86
4.5	Shortest paths (L^C)	18.04
4.6	Trips on a weekday	15.05
4.6	Trips on a weekend	16.32

(up to real time), and broader spatial and temporal coverage of city dwellers (Calabrese et al., 2013; Diao et al., 2016; González et al., 2008; Ratti et al., 2006; Reades et al., 2007). Furthermore, assuming that the smartphone applications have been implemented reliably, there is the potential to record a much more accurate representation of daily mobility than that captured by traditional methods such as travel diaries. However, these advantages are offset by the serious potential for bias in the data. Whereas focused surveys can be designed to be widely representative of the population, or at least reweighted afterwards, data collected through smartphone applications will almost certainly be a much poorer representation of society at large. Coupled with this is a general lack of reliable socio-demographic information to supplement the mobile data, which makes it difficult or impossible to reweight the sample. It is likely that the data used here are skewed to more affluent areas or individuals, as well as younger age groups. Whilst this does not invalidate the results—it is not unlikely that less affluent, older people for example will exhibit similar behaviors relating to route asymmetry—the results must be taken with the caveat that it is impossible to test their representativeness.

Whilst there appeared to be only a small difference between the asymmetry of weekday and weekend trips, it could be possible that there is a larger disparity in asymmetry due to the time of day of the trip. For example, people might be more likely to deviate on their route home (e.g., to go shopping) than in the morning on their way to work. This is an avenue that can be explored immediately in the future. More broadly, it would be interesting to explore the degree to which asymmetry is influenced by the underlying purpose of the journey. Therefore, an area of future work will be to attempt to estimate the trip purpose, and then begin to relate this to asymmetry. Initially this could be achieved by identifying important "anchor points" for each user (e.g., Malleson & Birkin, 2014) and then assigning these to likely functions (e.g., home, work, leisure, etc.). The subsequent highlighting of these travel behaviors may implicitly reveal factors about the built environment—amenities (Ewing et al., 2006), stores (Forsyth et al., 2008), social conditions (Jacobs, 1961), exposure to the elements (Cervero & Duncan, 2003), paths (Lynch, 1960), etc.—that may explain why certain route choices are made along one path versus another. Furthermore, it would be possible to examine the route data available here in the context of additional measures of environmental attractiveness—Quercia, Schifanella, and Aiello (2014), for example, use a crowd-sourcing mechanism to quantify "beautiful, quiet, and happy routes"—to better understand the impacts of these characteristics on route choice. Ultimately, quantifying the strength of these related factors will have implications for the creation and promotion of walkable communities and active transportation.

NOTES

[1] Data licenses require the name of the application to remain anonymous.

[2] It is theoretically possible that two cells will be adjacent because they share a corner, but in practice this has not occurred and so four neighbors can safely be used.

ORCID

Nick Malleson 🆔 http://orcid.org/0000-0002-6977-0615

REFERENCES

Agramunt, S., Meuleners, L., Chow, K. C., Ng, J. Q., & Morlet, N. (2016). A validation study comparing self-reported travel diaries and objective data obtained from in-vehicle monitoring devices in older drivers with bilateral cataract. *Accident Analysis & Prevention*, *106*, 492–497.

Ahmed, M., Karagiorgou, S., Pfoser, D., & Wenk, C. (2015). *Map construction algorithms*. Cham, Switzerland: Springer.

Barros, A. P., Martínez, L. M., & Viegas, J. M. (2015). A new approach to understand modal and pedestrians route in Portugal. *Transportation Research Procedia*, *10*, 860–869.

Bekhor, S., Ben-Akiva, M. E., & Ramming, M. S. (2006). Evaluation of choice set generation algorithms for route choice models. *Annals of Operations Research*, *144*(1), 235–247.

Ben-Akiva, M., Bergman, M. J., Daly, A. J., & Ramaswamy, R. (1984). Modeling inter-urban route choice behaviour. In *Proceedings of the Ninth International Symposium on Transportation and Traffic Theory* (pp. 299–330). Delft, the Netherlands.

Brockmann, D., Hufnagel, L., & Geisel, T. (2006). The scaling laws of human travel. *Nature*, *439*(7075), 462–465.

Calabrese, F., Diao, M., Di Lorenzo, G., Ferreira, J., & Ratti, C. (2013). Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. *Transportation Research: Part C, Emerging Technologies*, *26*, 301–313.

Cervero, R., & Duncan, M. (2003). Walking, bicycling, and urban landscapes: Evidence from the San Francisco Bay area. *American Journal of Public Health*, *93*(9), 1478–1483.

Chown, E., Kaplan, S., & Kortenkamp, D. (1995). Prototypes, Location, and Associative Networks (PLAN): Towards a unified theory of cognitive mapping. *Cognitive Science*, *19*(1), 1–51.

Cools, M., Moons, E., Creemers, L., & Wets, G. (2010). Changes in travel behavior in response to weather conditions: Do type of weather and trip purpose matter? *Transportation Research Record*, *2157*, 22–28.

Diao, M., Zhu, Y., Ferreira, J., & Ratti, C. (2016). Inferring individual daily activities from mobile phone traces: A Boston example. *Environment & Planning B*, *43*(5), 920–940.

Ewing, R., Handy, S., Brownson, R. C., Clemente, O., & Winston, E. (2006). Identifying and measuring urban design qualities related to walkability. *Journal of Physical Activity & Health*, *3*(Suppl. 1), S223–S240.

Forsyth, A., Hearst, M., Oakes, J. M., & Schmitz, K. H. (2008). Design and destinations: Factors influencing walking and total physical activity. *Urban Studies*, *45*(9), 1973–1996.

Garbrecht, D. (1970). Frequency distributions of pedestrians in a rectangular grid. *Journal of Transport Economics & Policy*, *4*(1), 66–88.

Gehl, J. (2011). *Life between buildings: Using public space*. Washington, DC: Island Press.

Glasgow, M. L., Rudra, C. B., Yoo, E.-H., Demirbas, M., Merriman, J., Nayak, P., . . . Mu, L. (2014). Using smartphones to collect time activity data for long-term personal-level air pollution exposure assessment. *Journal of Exposure Science & Environmental Epidemiology*, *26*, 354–362.

Golledge, R. G., Smith, T. R., Pellegrino, J. W., Doherty, S., & Marshall, S. P. (1985). A conceptual model and empirical analysis of children's acquisition of spatial knowledge. *Journal of Environmental Psychology*, *5*(2), 125–152.

González, M. C., Hidalgo, C. A., & Barabási, A.-L. (2008). Understanding individual human mobility patterns. *Nature*, *453*(7196), 779–782.

Hägerstrand, T. (1970). What about people in regional science? *Papers of the Regional Science Association*, *24*(1), 6–21.

Helbing, D. (1991). A mathematical model for the behavior of pedestrians. *Behavioral Science*, *36*(4), 298–310.

Helbing, D., Molnár, P., Farkas, I. J., & Bolay, K. (2001). Self-organizing pedestrian movement. *Environment & Planning B*, *28*(3), 361–383.

Hillier, B. (2015). *Space is the machine: A configurational theory of architecture*. Retrieved from http://spaceisthemachine.com/

Jacobs, J. (1961). *The death and life of great American cities*. New York, NY: Random House.

Lee, C., & Moudon, A. V. (2006). The 3ds+R: Quantifying land use and urban form correlates of walking. *Transportation Research: Part D, Transport & Environment*, *11*(3), 204–215.

Liu, L., Hou, A., Biderman, A., Ratti, C., & Chen, J. (2009). Understanding individual and collective mobility patterns from smart card records: A case study in Shenzhen. In *Proceedings of the 12th International IEEE Conference on Intelligent Transportation Systems* (pp. 1–6). St. Louis, MO: IEEE.

Lynch, K. (1960). *The image of the city*. Cambridge, MA: MIT Press.

Malleson, N., & Birkin, M. (2014). New insights into individual activity spaces using crowd-sourced big data. In *Proceedings of the 2014 ASE BigData/SocialCom/CyberSecurity Conference*. Palo Alto, CA: ASE.

Millonig, A., & Schechtner, K. (2007). Decision loads and route qualities for pedestrians – Key requirements for the design of pedestrian navigation services. In N. Waldau, P. Gattermann, H. Knoflacher, & M. Schreckenberg (Eds.), *Pedestrian and evacuation dynamics 2005* (pp. 109–118). Berlin, Germany: Springer.

Newson, P., & Krumm, J. (2009). Hidden Markov map matching through noise and sparseness. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (pp. 336–343). Seattle, WA: ACM.

Palmer, J. R. B., Espenshade, T. J., Bartumeus, F., Chung, C. Y., Ozgencil, N. E., & Li, K. (2013). New approaches to human mobility: Using mobile phones for demographic research. *Demography*, *50*(3), 1105–1128.

Phithakkitnukoon, S., & Ratti, C. (2011). Inferring asymmetry of inhabitant flow using call detail records. *Journal of Advances in Information Technology*, *2*(4), 239–249.

Quercia, D., Schifanella, R., & Aiello, L. M. (2014). The shortest path to happiness: Recommending beautiful, quiet, and happy routes in the city. In *Proceedings of the 25th ACM Conference on Hypertext and Social Media* (pp. 116–125). Santiago, Chile: ACM.

Ratti, C., Pulselli, R. M., Williams, S., & Frenchman, D. (2006). Mobile landscapes: Using location data from cell phones for urban analysis. *Environment & Planning B*, *33*(5), 727–748.

Reades, J., Calabrese, F., Sevtsuk, A., & Ratti, C. (2007). Cellular census: Explorations in urban data collection. *IEEE Pervasive Computing*, *6*(3), 30–38.

Sadalla, E. K., Burroughs, W. J., & Staplin, L. J. (1980). Reference points in spatial cognition. *Journal of Experimental Psychology: Human Learning & Memory*, *6*(5), 516–528.

Sadik-Khan, J., & Solomonow, S. (2017). *Streetfight: Handbook for an urban revolution*. New York, NY: Penguin Books.

Seaborn, C., Attanucci, J., & Wilson, N. (2009). Analyzing multimodal public transport journeys in London with smart card fare payment data. *Transportation Research Record*, *2121*(1), 55–62.

Seneviratne, P. N., & Morrall, J. F. (1985). Analysis of factors affecting the choice of route of pedestrians. *Transportation Planning & Technology*, *10*(2), 147–159.

Speck, J. (2013). *Walkable city: How downtown can save America, one step at a time*. New York, NY: North Point Press.

Streeter, L. A., Vitello, D., & Wonsiewicz, S. A. (1985). How to tell people where to go: Comparing navigational aids. *International Journal of Man-Machine Studies*, *22*(5), 549–562.

Toole, J. L., de Montjoye, Y.-A., González, M. C., & Pentland, A. (2015). Modeling and understanding intrinsic characteristics of human mobility. In B. Gonçalves & N. Perra (Eds.), *Social phenomena* (pp. 15–35). Cham, Switzerland: Springer.

Whyte, W. H. (2010). *The social life of small urban spaces*. New York, NY: Project for Public Spaces.

Wolf, J., Guensler, R., & Bachman, W. (2001). Elimination of the travel diary: Experiment to derive trip purpose from global positioning system travel data. *Transportation Research Record*, *1768*, 125–134.

Zacharias, J. (2001). Path choice and visual stimuli: Signs of human activity and architecture. *Journal of Environmental Psychology*, *21*(4), 341–352.