Scherrer et al. *EPJ Data Science* (2018) 7:19
https://doi.org/10.1140/epjds/s13688-018-0147-7

EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE

Open Access

Travelers or locals? Identifying meaningful sub-populations from human movement data in the absence of ground truth

Luca Scherrer[1]*, Martin Tomko[2], Peter Ranacher[1] and Robert Weibel[1]

*Correspondence:
lscherrer7@gmail.com
[1] Department of Geography, University of Zurich, Zurich, Switzerland
Full list of author information is available at the end of the article

Abstract

As users of mobile devices make phone calls, browse the web, or use an app, large volumes of data are routinely generated that are a potentially useful source for investigating human behavior in space. However, as such data are usually collected only as a by-product, they often lack stringent experimental design and ground truth, which makes interpretation and derivation of valid behavioral conclusions challenging. Here, we propose an unsupervised, data-driven approach to identify different user types based on high-resolution human movement data collected from a smartphone navigation app, in the absence of ground truth. We capture spatio-temporal footprints of users, characterized by meaningful summary statistics, which are then used in an unsupervised step to identify user types. Based on an extensive dataset of users of the mobile navigation app *Sygic* in Australia, we show how the proposed methodology allows to identify two distinct groups of users: 'travelers', visiting different areas with distinct, salient characteristics, and 'locals', covering shorter distances and revisiting many of their locations. We verify our approach by relating user types to space use: we find that travelers and locals prefer to visit distinct, different locations in the Australian cities Sydney and Melbourne, as suggested independently by other studies. Although we use high-resolution GPS data, the proposed methodology is potentially transferable to low-resolution movement data (e.g. Call Detail Records), since we rely only on summary statistics.

Keywords: Human mobility; Clustering; PCA; User characterization; Unsupervised learning; Movement patterns

1 Introduction

Today, a large part of data capturing human spatial mobility and behavior is being generated as byproducts of digital or online activities, for instance, during mobile phone use (Csáji et al. [9], Ahas et al. [1]) or as a by-product of taxi dispatching systems (Gong et al. [18]). Such data are often called *exhaust data* (Mayer-Schönberger and Cukier [34], Neef [35], George et al. [16]) and more specifically *exhaust human movement data* (EHMD), if they record human movement, such as in the examples given above. EMHD represent a source of potentially useful information about human behavior in space and time. As such EHMD are, however, not collected following stringent experimental design; they lack a suitable experimental foundation along with ground truth and demographic parameters of the population from which the data originate. Without controlling for population biases, it is thus difficult to interpret the findings gained from a particular data set and draw generalizable conclusions (Calabrese et al. [7], Zhao et al. [53]).

So, how can we reliably characterize subgroups of a moving population in the absence of verifiable ground truth data? In this article, we present a fully data-driven approach to identify distinct subgroups of moving populations based on EHMD collected from a mobile navigation app. We find that the movement behavior of people belonging to a population (sub)group is indicative of the group's space use. In other words, *how* individuals move (e.g. covering large distances and areas) shows us *where* they move (e.g. in or between certain city districts). This separation of behavior represented in relative spatial displacements on the one hand, from absolute spatial position on the other hand, allows us to evaluate the identified groups and supports intuitions about the group membership in the absence of ground truth data. To illustrate our approach, we use positioning data from a mobile navigation app for the study area of Australia. We find two primary, highly distinct groups of users that stand out by their behavior:

- 'travelers': users that move around extensively, visit different areas with distinct, salient characteristics and do not stay long at a specific location.
- 'local residents': users that move in a more constrained area, e.g. a city, usually cover only short distances and revisit many of their locations.

Moreover, we find that both groups prefer to visit distinct locations in the two most populous Australian cities, Sydney and Melbourne. Our methodology and findings can be used to inform nuanced urban population mixing models supporting spatio-temporal diffusion analysis for, e.g., epidemiological modeling, in the absence of detailed demographic information about the moving population. Using traditional authoritative data sources on demographics, such as census data, typically captures only the resident population. The proposed methodology is capable of distinguishing between locals, travelers and other demographic groups purely based on their movement characteristics. Assuming that many travelers are indeed tourists, the results of the proposed methodology can be used, for instance, to study flows and interactions between tourism precincts (Kelly [28]).

2 Related work

2.1 The value of EHMD

The recent substantial surge in human movement data being generated has led to an increased exploitation of data-driven approaches enabling to explore, summarize, and even predict the behavior captured by the movement data.

Humans use GPS devices (Pappalardo et al. [37]), log into Wi-Fi networks (Ren et al. [43]), use their mobile phones (Ahas et al. [1]), travel on subway systems (Lathia and Capra [29]) or take a taxi (Gong et al. [18]), all of which are activities that potentially generate exhaust data. Exhaust human movement data in particular share the following characteristics that make their analysis difficult:

EHMD are highly episodic: EHMD are only recorded when the user engages in a digital activity (Calabrese et al. [8]) or if a certain set of conditions are met during the broader data collection. Thus, EHMD often lack continuity (Phithakkitnukoon et al. [40]).

EHMD contain implicit semantics: In a trajectory, a journey from *home* to *work* is implicitly encoded as a sequence of spatio-temporal locations, but without the explicit semantic information about the origin and destination that would allow to interpret the intent of the trip. Attaching semantic information to EHMD requires sophisticated and careful processing (Calabrese et al. [7]).

EHMD are biased: Without controlling for population biases, it is difficult to interpret and draw generalizable conclusions from the findings gained from a particular data set (Zhao et al. [53]). Firstly, only specific groups may take part in digital activities (Zhang et al. [52], Birenboim and Shoval [4]). Secondly, not all movement is recorded, e.g. GPS navigation devices may only be used when traveling unknown routes and therefore EHMD do not capture a representative picture of the population behavior.

EHMD lack ground truth: as the intent behind the spatio-temporal behavior must be inferred from the data and the socio-demographic attributes of individuals are not controlled for during data collection (or are often not available at all), the correct labeling and interpretation of the data and analytical results are either intractable, or reliant on the skill of the analyst (Calabrese et al. [7]).

Yet, in earlier studies EHMD have supported important findings about human mobility: Humans move regularly, follow a reproducible pattern (Song et al. [46]) and return to a few significant locations (González et al. [19]). Humans who move similarly tend to have more intensive social connections, and share more interactions (Wang et al. [50]). Most importantly, EMHD have the potential to cover complete populations, or at least large samples of a population, as opposed to traditional data following an experimental design, which usually represent only small samples, typically counting a few dozens, and at most several hundreds or several thousands of users.

2.2 How to process EHMD

Data mining methods have the potential to efficiently discover hidden structures in data, thus supporting the discovery of patterns and clusters, or automatically assigning labels from a pre-defined set of classes to previously unseen data (Witten et al. [51], Han et al. [20]).

In all cases, data pre-processing fundamentally determines the success of data mining. Attribute reduction, via attribute subsetting and attribute construction (aka feature engineering) fundamentally impact on the results of the analysis (Witten et al. [51], Han et al. [20]). Attribute subsetting and dimensionality reduction lead to a more compact representation of the dataset while preserving its integrity. Attribute subsetting removes attributes without descriptive relevance. Dimensionality reduction recodes the data into a lower number of dimensions, constructing new features describing a set of the original attributes. Our approach, as outlined in Sect. 3, combines methods from feature extraction, dimensionality reduction and unsupervised learning on movement data.

2.3 Data mining with movement data

Data Mining has been successfully applied to the analysis of movement data. This includes, among others, the exploration of the dependencies between urban land use and space use (Pan et al. [36]), the classification of moving objects and mode of transport from the recorded characteristics of their movement, (Zheng et al. [54], Dodge et al. [13]), the understanding of the spatio-temporal and demographic patterns of human movement (Csáji et al. [9]), and the inference of environmental pollution (i.e., noise) based on large-scale crowd sensing (Zheng et al. [55]). A good overview of the techniques applied to feature extraction, characterization, and mining of movement data based on trajectories is available in a series of complementary papers of (Parent et al. [39], Lin and Hsu [32], Dodge, Laube, & Weibel [12]), focusing on linking movement data with trajectory semantics, on the computational aspects of trajectory processing, and on trajectory similarity, respectively.

These techniques are applicable in a wide range of domains, from refined algorithms supporting efficient carpooling (Trasarti et al. [49]), through characterization of commuting patterns (Csáji et al. [9]), to the understanding of the nature of movement (Lee et al. [30], Pappalardo et al. [38]). In this paper, we explore how characteristic parameters of a user's movement behavior are manifested in their space use, thus combining two distinct perspectives on movement analysis to gain insights about the typology of the user.

2.4 Travelers and locals: characterization of users by movement behavior

Understanding the spectrum of the urban population and their space use is of fundamental importance for the management of cities, their transportation infrastructure, the provision of adequate services and governance, and the maintenance of public safety. Travelers and locals have distinct needs in an environment. In one of the first studies of EHMD for urban movement behavior analysis, Girardin et al. [17] studied the digital footprints captured from image databases (Flickr) and collected from aggregated call record data (e.g. Call Detail Records, CDR). Primarily focusing on the visualization of the users' behavior, they explored so called *desire lines* constructed from digital traces, to contrast space use by international tourists from the USA and Italians in the city of Rome. The information about the country of origin of the users was, crucially, available from the mobile phone subscription. They suggest that this kind of data complements traditional surveys and data collection about tourism behavior. While call record data are difficult to access, annotated Flickr photographs remain a popular, and accessible source of information about tourist space use worldwide (Kádár and Gede [26]). These approaches, however, rely on the ability to distinguish locals and (usually only international) tourists based on a controllable piece of information. Despite their importance to the local economy, domestic tourists and locals are often not studied or included in studies focused on tourism (Hede and Hall [23]).

On the opposite end of the level of detail, Asakura and Iryo [2] studied highly detailed patterns of user movement captured by GPS. Their clustering-based method enabled to identify tourists with similar trajectories using an index of trajectory topology. Yet, this approach was based on a controlled study of pedestrians in a highly constrained area of interest (approx. 300 m × 300 m). Here, we capture numerous characteristics of user trajectories not constrained to a specific mode of transport or area, to improve the classification across coarser behavioral groups, rather than identifying the specific space use in a given city, by (known) user type. As such, our contribution provides a method applicable in situations where ground truth about the user type is not available. More detailed studies can then be targeted at the identified subpopulation of the tracked population, possibly within a constrained spatial area to analyze detailed patterns of space-use of, for example, travelers or locals (Edwards and Griffin [14]).

3 Methodology

This section describes the proposed methodology to infer user types from EHMD, in the absence of ground truth. We describe the data used to illustrate our approach and introduce the steps of the methodology one by one. Hence, each individual step of the proposed methodology is not only introduced and described, it is also directly applied to the EHMD data set that we use as a case study.

3.1 Data

Movement data of mobile users were provided by *Sygic*, a mobile app service for global navigation assistance with over 100 Million users (Sygic [47]). Users run the *Sygic* app when requiring additional information relating to their travel. This includes navigation instructions if they are not familiar with their environment and in need of assistance to follow a route; assistance information for wayfinders locating an unknown place in an otherwise familiar environment; and additional traffic information for motorists about congestions, accidents, or speed cameras. It is, therefore, reasonable to assume that the *Sygic* app users will include a mix of different types, such as users traveling longer distances, as well as users in more local transportation settings, such as commuters or delivery drivers, as different user groups are serviced by at least some of the functions provided by the *Sygic* app. *Sygic* does not collect information about the users. Accordingly, public transport users and pedestrians might use the app as well. The navigation aid is, however, clearly targeted at drivers of motorized vehicles. We therefore argue that only an insignificant percentage of the app users are non-motorists, and these would not change the patterns detected.

Anonymized user data were available for this study, recorded in Australia between January 1 and January 31, 2016, capturing the movement of 71,207 unique users. Accordingly, no data from the *Sygic* Travel app was used, as it was only acquired by Sygic in April 2016. Australia offers an ideal test area due to its isolation and lack of land border crossings, thus representing an encapsulated system.

The data comprise of GPS tuples with position (latitude/longitude) and timestamp. Each tuple has a unique ID, which relates to a specific user. Tuples are recorded at a five second interval. Data are not logged locally (i.e. they are stored on *Sygic* servers) and are not recorded when the app is not actively used or is outside the range of mobile coverage.

3.2 Overview

The overview of our user type inference methodology from spatio-temporal footprints is shown in Fig. 1. We define a *spatio-temporal footprint* as the aggregated movement (equaling the entirety of a user's recorded trajectories) as seen in the raw positioning data (in our case GPS). Accordingly, it can be seen as a proxy of the user's spatio-temporal behavior over time. A *user type* then denotes a certain group of users with a similar behavior. An example of a user type is a traveler, who visits different areas and only stays for a brief period of time at a single location. It is important to note that user types are derived from the data and not defined *a priori*.

Figure 1 summarizes the main steps of our approach, described in detail in the following sections. After *data cleaning*, we compute a set of meaningful behavioral spatio-temporal features for each user with a spatio-temporal footprint (*feature extraction*). Again, these features do not relate to a specific *absolute* location in space and time, but rather describe the *relative* movement of a user, for example the average extent of the area a user has covered in a single day. Accordingly, users roaming in two different cities may have similar characteristics of their footprints, although they do not visit the same places. We then perform a principal component analysis (PCA) to single out the most informative combi-

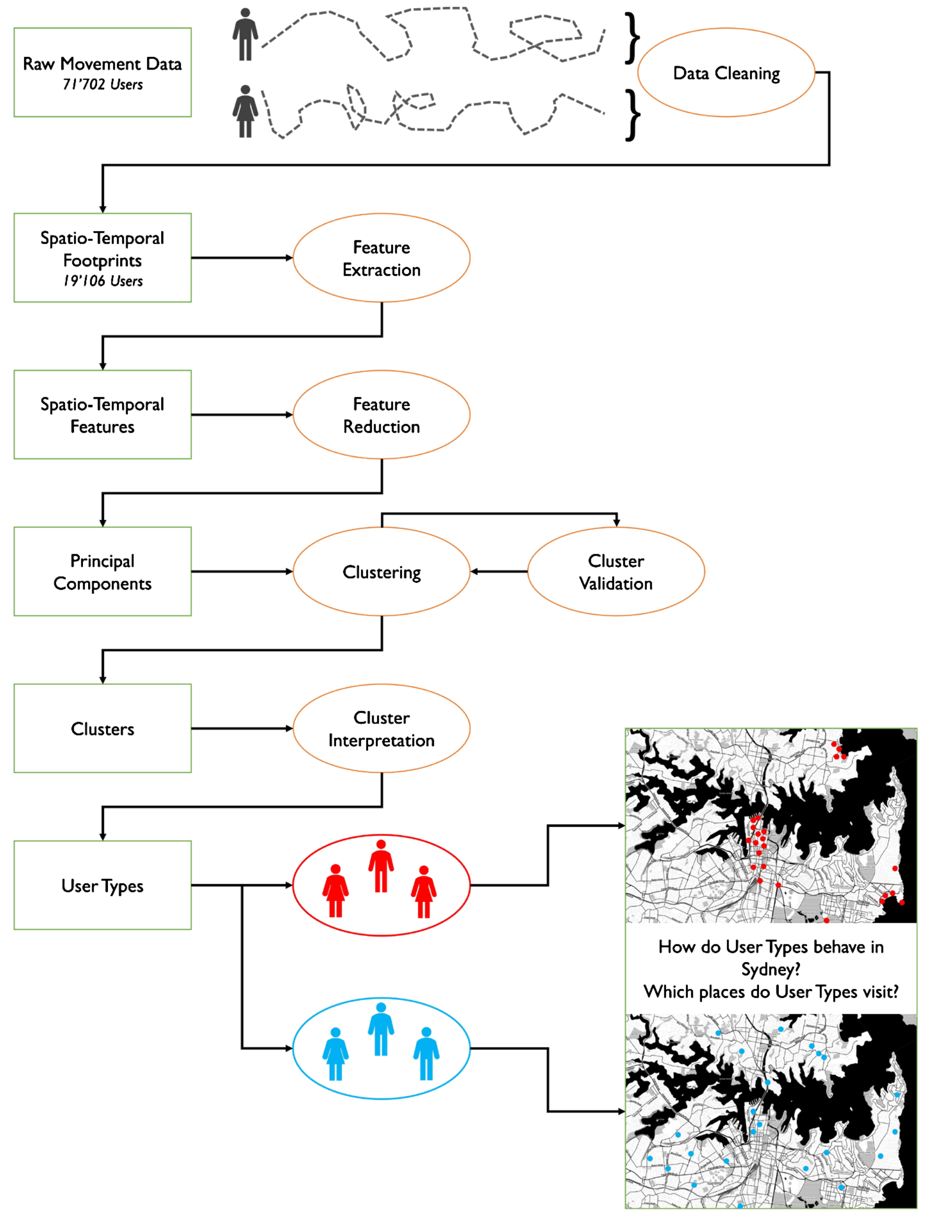

Figure 1 Inferring User Types from raw movement data

nation of features (*feature reduction*). According to the principal components, we cluster the users into distinct groups (*clustering*), as a form of unsupervised learning. We interpret these groups of user types (*cluster interpretation*) and explore their behavior in the cities of Melbourne and Sydney, Australia, thus relating the characteristics of their movement to their spatial setting.

In this proposed methodology, the feature extraction and dimensionality reduction steps make explicit human behavior that is implicitly hidden in exhaust human movement data, while the unsupervised learning by clustering allows us to draw inferences and interpretation without ground truth.

3.3 Data cleaning

We cleaned the data by removing users with almost no movement (<300 m) during a single day; users with abnormal speed values (>180 km/h); and lastly users that did not run the app for at least 5 days out of the 31 days for which we had data. After the cleaning, 19,106 users—from originally 71,702—remained for further analysis.

3.4 Feature extraction

For each user we defined 32 features, which we derived from the spatio-temporal footprints (see Table 1). A feature is a meaningful summary statistic about a specific aspect of a user's movement. Features are solely based on relative spatial, temporal, and spatio-temporal characteristics and are divided into five types: *Temporal Activity* (6 features), *Spatial Distance* (10), *Spatial Area* (5), *Spatial Variability* (4), and *Spatio-Temporal Dynamics* (7).

Temporal Activity captures how often, how regularly and for how long a user engages with the app. We expect to distinguish active and occasional users. *Spatial Area* comprises statistics about the area in which a user roams, its shape, size and extent. We expect to distinguish users with different areal patterns, i.e. users who visit small, compact regions and those who roam in the entire country. *Spatial Distance* comprises statistics about the spatial path of a user, its stages and its distance covered. We expect to distinguish users with mostly short trips and those with mostly long journeys. *Spatial Variability* comprises features relating to the variability of a user's location in space. We expect to distinguish spatially stable users, roaming between a few distinct clusters, and spatially volatile users, visiting many locations dispersed in space. Finally, *Spatio-Temporal Dynamics* relate to the behavior of a user in space and time. We expect to distinguish users with continuous and variable movement patterns, the former moving uniformly, and the latter changing their mobility dynamics over time. The 32 features serve as a basis for dividing the individual users into distinct groups and for inferring user types. As such, they capture characteristics shared by a *type* of users, whether they are in Sydney, Melbourne, or elsewhere in the area covered by the trajectory data (Australia, in our case).

3.5 Iterative feature reduction and clustering

The 32 computed features may not be equally important for distinguishing distinct groups of users. Principal component analysis (PCA) (Bro and Smilde [5]) is a data reduction technique enabling to reduce a dataset described by data vectors with n attributes to a dataset described by k n-dimensional vectors capturing the bulk of the variation in the dataset (James et al. [24]).

A subsequent cluster analysis of the principal components (PC) enables to identify cohesive, meaningful groupings in the data. There are numerous methods for clustering multidimensional data (Kaufman and Rousseeuw [27]), yielding distinct results of different plausibility. A suitable method should lead to cohesive, useful and interpretable clusters (James et al. [24]). The performance of a particular clustering method depends on the problem at hand, as well as on the expected or desired number of clusters and the explanatory power of the input data (here, the number of principal components used to capture the variation in the data). These three parameters are mutually dependent and cannot be decided on in isolation. We have therefore designed an iterative data-driven approach to identify an optimal clustering method with an optimal number of clusters and an optimal number of principal components to identify user types in the data.

Table 1 The initial 32 features derived from the spatio-temporal footprint of each user

	Feature name	Clarification and description	Feature type
1	Active days	The number of days the user has used the app	Temporal Activity
2	Consecutive days	The highest number of consecutive days the user has used the app	Temporal Activity
3	Weekdays	The number of distinct weekdays the user has used the app	Temporal Activity
4	Period	The number of days between the first and last usage	Temporal Activity
5	Total time	The total amount of time the app has been running [s]	Temporal Activity
6	Variation of time stamp	The standard deviation of all time stamps [s]	Temporal Activity
7	Total distance	The total distance the user has covered [m]	Spatial Distance
8	Maximum distance	The distance between two most distant points a user has visited [m]	Spatial Distance
9	Daily distance	The average distance covered in a day [m]	Spatial Distance
10	Variation of daily distance	The standard deviation of the average distances per day [m]	Spatial Distance
11	Daily centroid distance	The average distance of two consecutive[a] daily centroids [m] The centroid is the centroid of the daily concave hull	Spatial Distance
12	Variation of daily centroid distance	The standard deviation of the distance between two consecutive daily centroids [m]	Spatial Distance
13	Distance to centroid	The average distance between the daily centroid and the overall centroid [m]	Spatial Distance
14	Variation of distance to centroid	The standard deviation of the distance between the daily centroid and the overall centroid [m]	Spatial Distance
15	Average step length	The average distance covered in a move segment[b] [m]	Spatial Distance
16	Standard deviation of step length	The standard deviation of the distances covered in move segments [m]	Spatial Distance
17	Area	The total area the user has covered [m^2][c]	Spatial Area
18	Circumference	The circumference of the total area the user has covered [m]	Spatial Area
19	Complexity	The complexity of the total area (area/circumference)	Spatial Area
20	Compactness	The compactness of the total area [$4 * \text{area}/\pi *$ maximum distance squared]	Spatial Area
21	Daily area	The average area of the daily areas covered [m^2]	Spatial Area
22	Variation of daily area	The standard deviation of the daily areas covered [m^2]	Spatial Variability
23	Overlap	The average percent of overlap of two consecutive daily areas covered [%]	Spatial Variability
24	Variation of overlap	The standard deviation of the percentage of overlap of two consecutive daily areas covered [%]	Spatial Variability
25	Spatial clusters	The number of clusters of start, stop or end points[d]	Spatial Variability
26	Number of moves	The absolute number of move segments	Spatio-Temporal Dynamics
27	Average speed	The average speed in the move segments [m/s]	Spatio-Temporal Dynamics
28	Standard deviation of speed	The standard deviation of the speed in the move segments [m/s]	Spatio-Temporal Dynamics
29	Number of stops	The total number of stops	Spatio-Temporal Dynamics

Table 1 (*Continued*)

	Feature name	Clarification and description	Feature type
30	Total stop duration	The total duration of all stops [s]	Spatio-Temporal Dynamics
31	Stop duration	The average duration of a stop [s]	Spatio-Temporal Dynamics
32	Variation of stop duration	The standard deviation of the stops [s]	Spatio-Temporal Dynamics

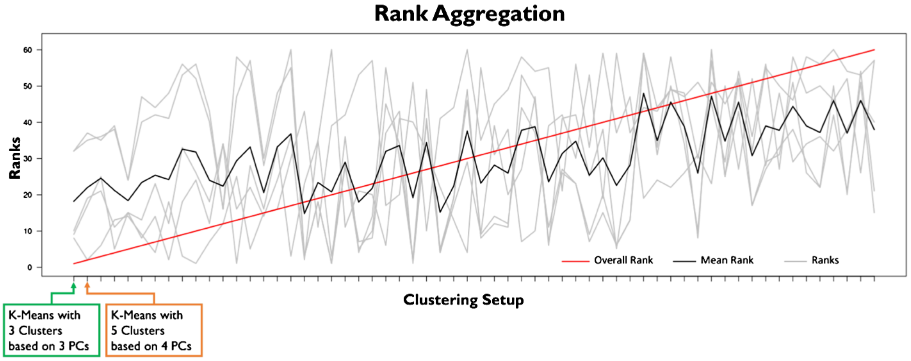

Figure 2 Rank aggregation of different clustering setups

Four clustering algorithms—DIANA, CLARA, AGNES (Kaufman and Rousseeuw [27]) and *k*-means (MacQueen [33]), five different sets of clusters (two to six clusters), and three sets of principal components (three, five and six PCs)—have been tested, leading to a total of 60 different clustering setups. The results of these setups were tested using five distinct statistical tests per setup. The tests comprise the silhouette width (Rousseeuw [44]), the gap statistic (Tibshirani et al. [48]) and three stability measures: average proportion of non-overlap, average distance, as well as average distance between means (Datta and Datta [11]). The results of the tests have been aggregated to decide on the optimal combination of parameters.

3.6 Finding an optimal clustering approach

Rank aggregation of all test results (Pihur et al. [41]) enables to find a consensus between the ranked lists from the five test statistics applied to the clustering setups. It generates an overall ranking that shows the highest consistency with all individual ranked lists. When computing the final rank aggregation of the test results, the five clustering quality measures were weighted unequally. The silhouette width and the gap statistic list were each weighted one, whereas the three stability measures together were weighted one third each, since they are similar in nature and are expected to yield similar results. Kendall's tau distance was applied to measure the distances between the ranked lists.

The rank aggregation in Fig. 2 shows the ranks of all setups for the five different clustering quality measures (light grey lines), the overall mean rank for each setup (dark grey line) and the weighted mean (red line). Accordingly, the best setup (*k*-means with three clusters and three PCs—green box) can be found on the left side of Fig. 2.

Statistical properties are, however, only one way of evaluating the goodness of clustering. It is also important to produce useful and interpretable clusters (Brock et al. [6], James et al.

[24]), which in our case would lead to meaningful user types. A final inspection step of the two best-ranked approaches performing at par led to the selection of the second ranked setup, a k-means clustering with 5 clusters based on 4 principal components (marked with the orange box in Fig. 2).

3.7 Interpretation

The selected clustering approach yields five groups of users with distinct spatio-temporal footprints. Labeling of these five groups is challenging, as it is based solely on the characteristics of the clusters in the absence of ground truth. Therefore, we first describe the clusters (see also Fig. 3) and only then label those with the most salient behavior.

- Cluster 1 (4451 unique users): Cluster members are inactive. They turn on the app only sporadically. Thus, their recorded movement is episodic and erratic. Moreover, users in Cluster 1 move slowly and stop frequently.
- Cluster 2 (3343 unique users): Cluster members are active, cover large distances and visit many different locations in Australia.
- Cluster 3 (5861 unique users): Cluster members are highly active, but only roam in relatively confined areas.
- Cluster 4 (1079 unique users): Similarly to Cluster 1, cluster members are inactive and only use the app sporadically. However, in contrast to Cluster 1, they move fast and stop only infrequently.

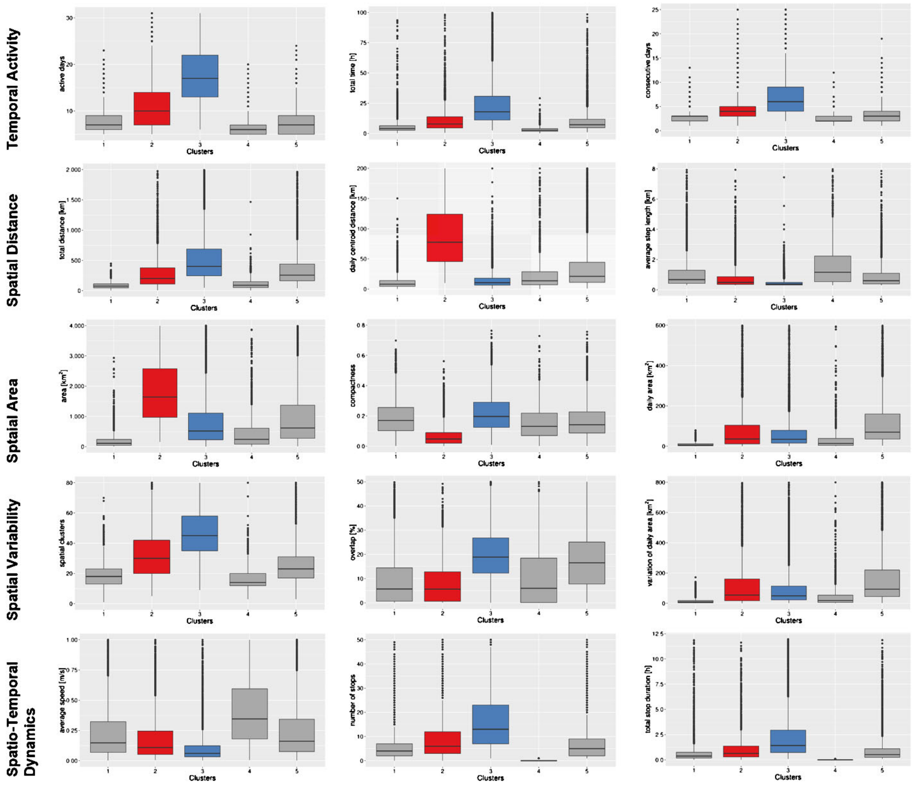

Figure 3 Some selected features for travelers (red) and locals (blue). Clusters 1, 4 and 5 (gray) are not interpreted in detail (see text)

– Cluster 5 (4372 unique users): Cluster members are active. On the one hand they show a behavior that lies *between* that found in Cluster 2 and Cluster 3, for example with respect to the distance travelled, the total area visited, or the compactness. On the other hand, they behave *similarly* to Clusters 2 and 3, e.g. with respect to the number of spatial clusters[e] (Cluster 2), and with respect to overlap (Cluster 3).

Users in Clusters 1 and 4 are only active during slightly more than the threshold of 5 days (Fig. 3). The recorded movement is short and episodic and, therefore, difficult to interpret. The clusters reflect how users engage with the app, rather than how they move. Users only turn on the app in situations where they need guidance and turn it off afterwards. Clusters 2 and 3 show the most salient and contrasting behavioral pattern. We label users in Cluster 2 'travelers', since they move between different locations in Australia covering large distances, and we label users in Cluster 3 'locals', since they mainly roam in a restricted area (for details see section below). Cluster 5 shows a non-salient behavior which is a mixture of the behavior in Cluster 2 and Cluster 3.

We only label and interpret in detail the two most salient Clusters 2 and 3. We argue that this is an important aspect of unsupervised learning: Although five clusters best capture the variance in the spatio-temporal footprints, an analyst should not expect that all clusters are semantically meaningful (Clusters 1 and 4) or that all clusters show salient behavior that can be interpreted in a straight-forward way (Cluster 5). The following section gives a detailed motivation and explanation for the labeling.

3.7.1 Travelers

Cluster 2 (3343 unique users) comprises of relatively active users, who run the app for 10 days on average. Their movement expands over a large area (1600 km^2), implying a low compactness (0.04). Users shift their daily centroid between consecutive days. This can be concluded from a large daily centroid distance (150 km), a large distance between the daily and the overall centroid (80 km) and a large maximum distance (480 km). Moreover, the movement over consecutive days hardly ever overlaps (6%). The overall distance covered (200 km) and the daily distance covered (20 km) are about average, which indicates steady but not excessive motion. The users move at average speed, their step length is slightly above average (480 m). They stop often (12 stops per trip) and head to numerous different destinations, which results in many spatial clusters of significant locations (30).

We label users in Cluster 2 as *travelers*: they change their center of activity, visit many different places and cover large parts of the country. Surprisingly, the total distance traveled in the cluster is not particularly large. We have two possible interpretations for this behavior:

– Users are familiar with some locations on their overall route, but not with all. Thus, they turn on the navigation app only sporadically.
– Users visit different regions in Australia, which they reach by other means of transport (e.g. planes). For these travels they do not need the navigation app. This behavior could be typical for tourists (in particular, domestic tourists but also business travelers). The large distances between major cities in Australia cause this to be a typical tourist behavior.

3.7.2 Locals

Cluster 3 (5861 unique users) comprises users who are highly active and run the app on median a total of 15 days (see also Fig. 3). The users have many spatial clusters (42) while roaming in a relatively small (median of 500 km^2) and compact (0.2) area. Furthermore, they are relatively stable with respect to their location: the daily overlap (20%) is high, the daily centroid distance (12 km) and the distance to the centroid (15 km) are small, and so is the maximum distance covered (60 km). Moreover, the users move slowly and in small steps (420 m).

We argue that users in Cluster 3 are *locals*, potentially roaming in larger cities and potentially using the app when commuting. The users stay in a relatively compact and small area and revisit many of their locations. The distances between consecutive stops are small and the average speed of the users is the lowest of all clusters. This is—potentially—indicative of driving in an urban environment with dense traffic.

4 Spatio-temporal behavior of user types

In this section we show *when* and *where* the two user types—the travelers and the locals—move in the cities of Sydney and Melbourne. First, we define significant locations (SL) for each user. An SL is a point in space and time that is either a start or an end point of a user's movement, or the first point of a significant stop segment. We therefore defined a stop segment as a series of GPS tuples with almost no movement (sum of covered distances shorter than 10 m) in a window of at least five minutes.

We expect a significant location to have a special significance to the individual user, i.e. it is a place that the user intends to visit rather than just passing by. However, we cannot fully rule out the possibility that a SL is simply a point where the user turned off the navigation app.

We now explore the distribution of significant locations in the cities of Sydney and Melbourne, both in space and time for both travelers and locals as the most significant groups of users identified.

4.1 Aggregated temporal patterns

The temporal behavior of travelers and locals in Sydney and Melbourne was analyzed for daily and weekly temporal distribution and periodicity.

Four weeks of data (28 days) were aggregated into non-overlapping six-hour windows (see Fig. 4). The absolute number of SL per user type was then standardized (yielding z-scores) to obtain time series that show the same variance. This allows us to compare the temporal trend of the two user types despite their different absolute number of SL (Fig. 4).

Table 2 shows the summary statistics for hourly and daily SL in Sydney and Melbourne, for both travelers and locals. As expected, locals are on average more present (Sydney: 1742 SL per hour/Melbourne: 1895 SL) than travelers (Sydney: 303 SL per hour/Melbourne: 323 SL). The hourly coefficient of variation (CV) is about the same for locals and travelers (∼0.6), whereas the daily CV is much higher for locals (0.12). In short, local app users visit more locations than travelers. At the same time their presence varies more from day to day.

We can further explore this pattern in Fig. 4. Travelers remain rather stable over the entire week, with two peaks in the morning and afternoon and one trough during the night. Locals have a behavior similar to that on weekdays (Monday to Friday). On weekends (Saturday and Sunday), however, they show only one daily peak, which is also less pronounced. For locals the two peaks during weekdays are of almost equal size, whereas for travelers the afternoon peak is slightly higher. This is also reflected in the travelers' slightly higher daily CV (see Table 2).

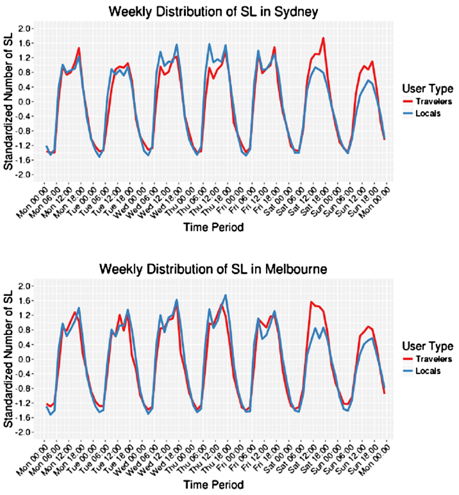

Figure 4 Standardized weekly distribution of significant locations for travelers (red line) and locals (blue line) in Sydney and Melbourne

Table 2 Summary statistics for hourly and daily SL in Sydney

		Travelers			Locals		
		Mean	Std. dev.	CV	Mean	Std. dev.	CV
Sydney	Hourly	303.06	202.30	0.67	1741.80	1035.43	0.59
	Daily	3636.71	238.96	0.06	20,901.57	2607.68	0.12
Melbourne	Hourly	323.35	227.23	0.70	1895.36	1089.02	0.57
	Daily	3880.14	284.43	0.07	22,744.29	2792.68	0.12

4.2 Aggregated spatial patterns

In a second step, we analyze the spatial behavior of travelers and locals in Sydney and Melbourne. We report their relative distribution, their location quotient, and their connectivity. We aggregated the significant locations per statistical area. We used the Statistical Areas Level 2 (SA2) proposed by the Australian Statistical Geography Standard (ASGS). SA2 areas are the second smallest statistical unit in Australia and represent "a community that interacts together socially and economically" (Australian Bureau of Statistics [3]). Furthermore, given that many travelers are potentially tourists, the spatial granularity of SA2 areas is roughly equal to typical 'tourism precincts' (Hayllar and Griffin [21]).

In the following visualizations, we highlight four special SA2s in the Greater Sydney Area and four SA2s in the Greater Melbourne Area, respectively. For Sydney, these are the Airport (A), the City Center (C, including *"The Rocks"* and the central business district) and two of Sydney's most famous beaches: Bondi Beach (B) and Manly Beach (M). For Melbourne, these are the Airport (A), the City Center (C) and the most famous beach area, St Kilda (S). These locations have also been identified as major tourism precincts in related empirical studies (Edwards et al. [15]).

4.2.1 Relative spatial distribution

For each user type we compute the percentage of significant locations per SA2, with a lower bound threshold of ≥12 SL/SA2 (a natural break in the distribution of SLs per area). This yields the relative distribution of travelers and locals in the two cities (Sydney: Fig. 5; Melbourne: Fig. 6).

As would be expected, locals are the most represented group in all areas of both cities. In Sydney, they especially dominate in the south and west of the city center, whereas in

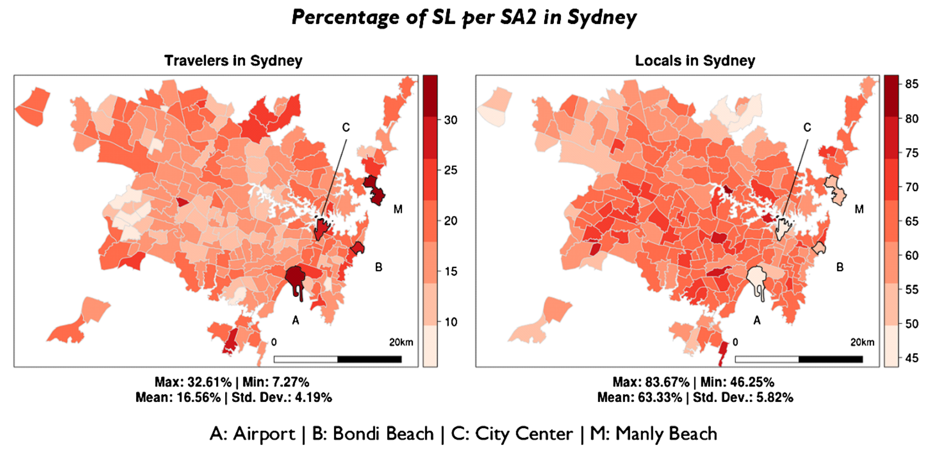
Percentage of SL per SA2 in Sydney
Travelers in Sydney
Locals in Sydney
C
M
B
A
0
20km
30
25
20
15
10
85
80
75
70
65
60
55
50
45
Max: 32.61% | Min: 7.27%
Mean: 16.56% | Std. Dev.: 4.19%
Max: 83.67% | Min: 46.25%
Mean: 63.33% | Std. Dev.: 5.82%
A: Airport | B: Bondi Beach | C: City Center | M: Manly Beach

Figure 5 Percentage of significant locations per SA2 in Sydney

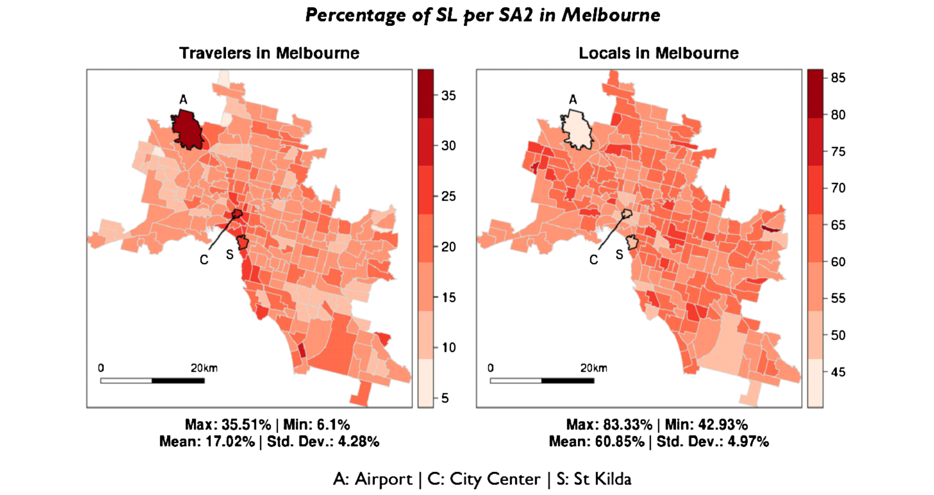
Percentage of SL per SA2 in Melbourne
Travelers in Melbourne
Locals in Melbourne
A
C
S
0
20km
35
30
25
20
15
10
5
85
80
75
70
65
60
55
50
45
Max: 35.51% | Min: 6.1%
Mean: 17.02% | Std. Dev.: 4.28%
Max: 83.33% | Min: 42.93%
Mean: 60.85% | Std. Dev.: 4.97%
A: Airport | C: City Center | S: St Kilda

Figure 6 Percentage of significant locations per SA2 in Melbourne

Melbourne, the areas west, north, and east of the city center are dominated by the locals. In both cities, travelers are most present in the airports, the city centers and the beaches (Manly and Bondi Beach in Sydney, St Kilda in Melbourne; >30%). Conversely, these are the SA2s with relatively least locals (<45%). Locals and travelers exhibit a similar variation over space. The standard deviation is around 5% for both groups.

4.2.2 Location quotient

In a next step we compute the *location quotient* (LQ) to identify SA2s with a non-standard visiting pattern. The location quotient compares the local density of a phenomenon in an area (one specific SA2) to the overall density of that phenomenon in a reference area (the whole Greater Area of Sydney/Melbourne) (Reades et al. [42], Jiang et al. [25]). A user type has a high LQ in an area where it is relatively overrepresented, and a low LQ where it is underrepresented. For each SA2 and each user type we compute the LQ as the relative difference between the SL observed for a particular SA2 and the expected SL, where the

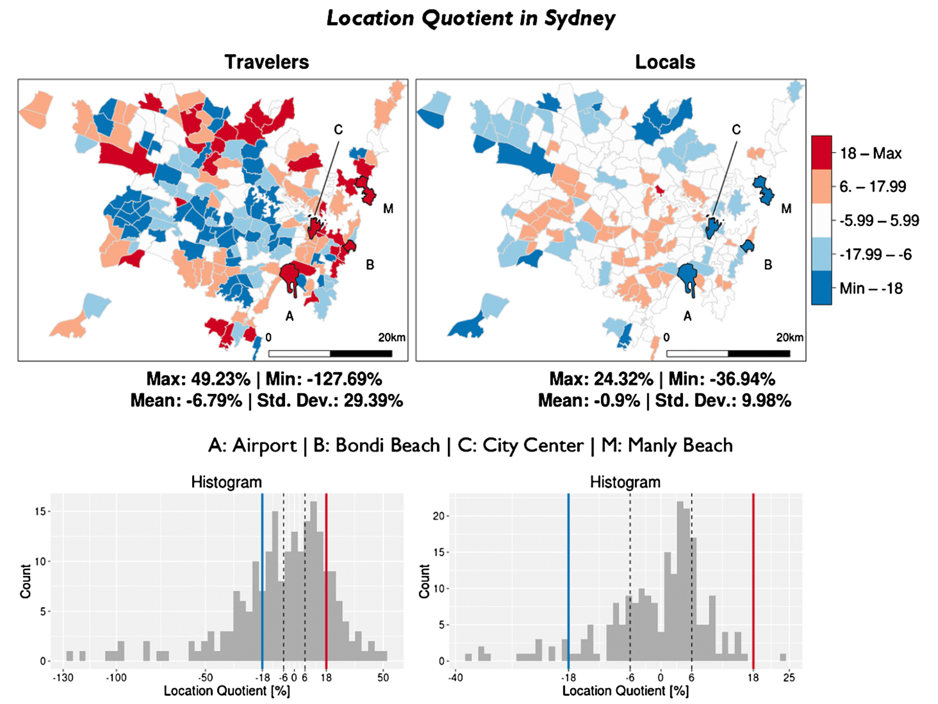
Location Quotient in Sydney
Travelers
Locals
C
M
B
A
0
20km
18 – Max
6. – 17.99
-5.99 – 5.99
-17.99 – -6
Min – -18
Max: 49.23% | Min: -127.69%
Mean: -6.79% | Std. Dev.: 29.39%
Max: 24.32% | Min: -36.94%
Mean: -0.9% | Std. Dev.: 9.98%
A: Airport | B: Bondi Beach | C: City Center | M: Manly Beach
Histogram
Count
Location Quotient [%]
-130
-100
-50
-18
-6
0
6
18
50
-40
-18
-6
0
6
18
25
0
5
10
15
20

Figure 7 Location quotients for each SA2 per user type in Sydney (top) and histograms of location quotients (bottom)

latter is the mean of all SL over the entire Greater Sydney Area or Greater Melbourne Area, respectively.

$$LQ = \frac{\text{observed SL} - \text{expected SL}}{\text{observed SL}} * 100.$$

For example, if the expected SL of travelers in Greater Sydney is 30%, but 50% of all observed SL in Manly Beach belong to travelers, the LQ for travelers in Manly Beach is +40%. Hence, travelers are relatively overrepresented in Manly Beach. In other words, when a user type is over- or underrepresented they may not be the most frequent originator of SLs in an area, but their SL frequency deviates the most from the expected mean value.

Figure 7 (Sydney) and Fig. 8 (Melbourne) show that travelers are overrepresented in the city centers, the airports and the areas around the beaches. Conversely, locals are underrepresented in these areas. In Sydney, locals have a high LQ for parts of western and southern Sydney, whereas in Melbourne, locals have a high LQ in the northwestern and southeastern areas. Travelers have more areas in both cities where they are either under- or overrepresented (maximum: 49% and 52%; minimum: –128% and –180%) compared to locals (maximum: 24% and 26%; minimum –37% and –42%).

4.2.3 Connectivity

Finally, we compute the connectivity between individual SA2s for both travelers and locals. *Connectivity* is defined here as the number of trips between an origin (O_i) and a destination (D_j), which in our case are two distinct SA2s in the Greater Sydney Area or Greater Melbourne Area, respectively. A trip is generated by a user traveling from O_i to D_j, which

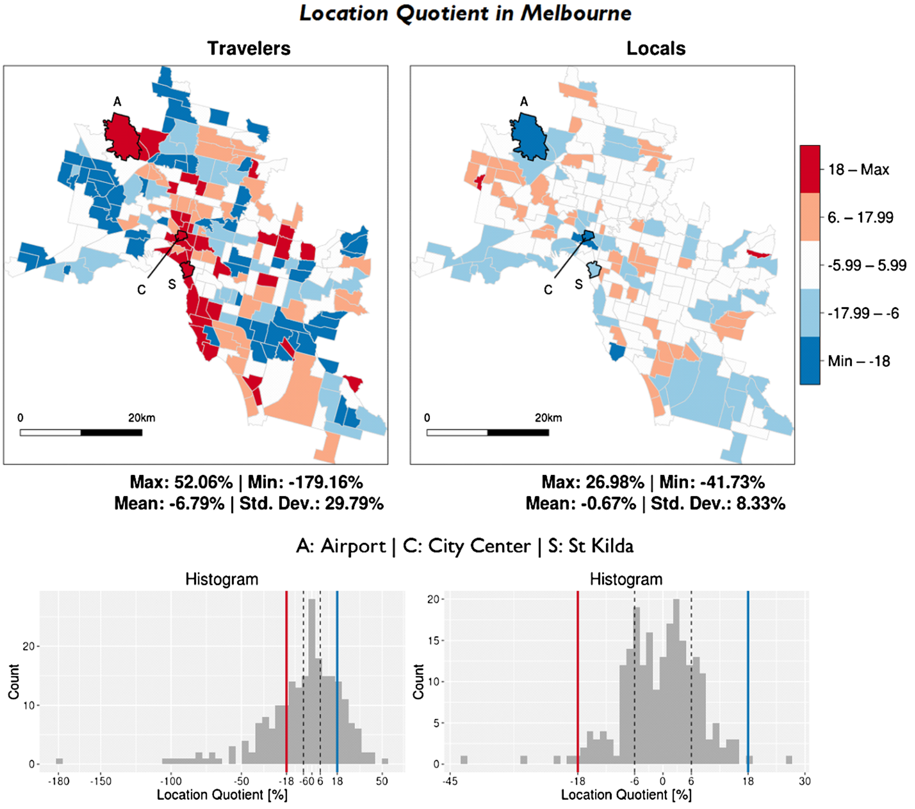

Figure 8 Location quotients for each SA2 per user type in Melbourne (top) and histograms of location quotients (bottom)

implies that the user has a first significant location in O_i and a consecutive one in D_j. We compute the OD matrix for all pairs of SA2s. If the trip is not entirely within the Greater Sydney Area or Greater Melbourne Area, respectively, it is not considered. Moreover, we only count one distinct trip between each pair of SA2s per user. This removes the potential bias of very active users who might, for example, commute daily between two areas.

Our results show that the two user types not only exhibit a distinct spatial and temporal pattern, but their trips also connect different areas in and around the two cities studied. In Fig. 9 (Sydney) and Fig. 11 (Melbourne), respectively, locals have a fine web of connections between many different areas, whereas travelers have a sparser web and move along fewer, distinct axes. These axes run along the main routes, mostly leading to the city center. Although locals also use these axes, these are less pronounced. Note the legend in Fig. 9 (also Fig. 10 and Fig. 11), which accounts for the different number of total trips for travelers and locals. The histograms at the bottom of Fig. 9 and Fig. 11 confirms this pattern. Travelers do not make trips between many of the SA2 pairs, hence the sparse web. There are a few positive outliers—the main axes, which can also be found for the locals. However, the locals' trips are less skewed towards low values, which implies a dense web of connections.

Figure 10 shows an enlarged view of Fig. 9 for the city center of Sydney. Along the main axes (city center, airport, Bondi and Manly Beach) both user types have a high number of trips. Apart from these, travelers have fewer connections and a sparser web.

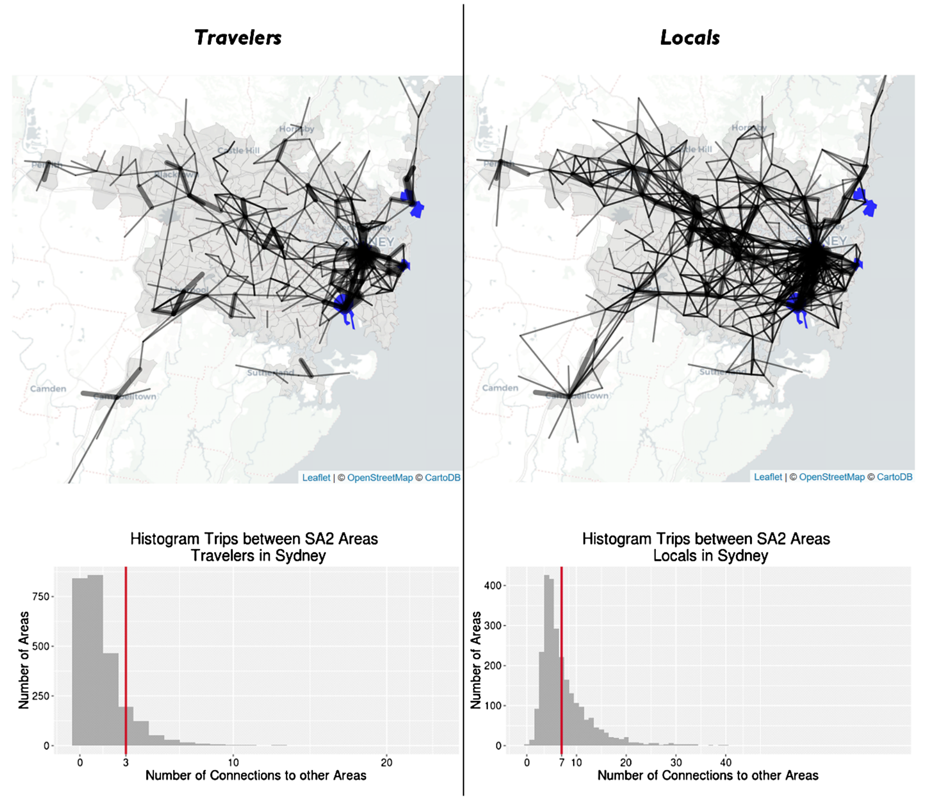

Figure 9 Top: The connectivity of SA2s in Sydney, the line width corresponds to the number of trips between the areas. Blue color indicates the airport, the city center, Bondi Beach, and Manly Beach. Bottom: Respective histograms of trips

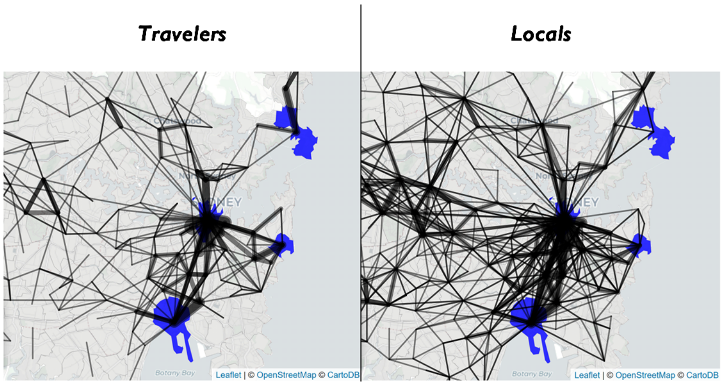

Figure 10 The connectivity of SA2s in Sydney, zoomed-in to the city center. Blue color indicates the airport, the city center, Bondi Beach, and Manly Beach

5 Conclusion

As EHMD become increasingly available, possibilities to extract meaningful information about human behavior emerge. Yet, the lack of ground truth data limits the applicability of EHMD to support decision making, governance and policy. Here, we presented an approach that enables to cross-reference independent features of recorded movement data

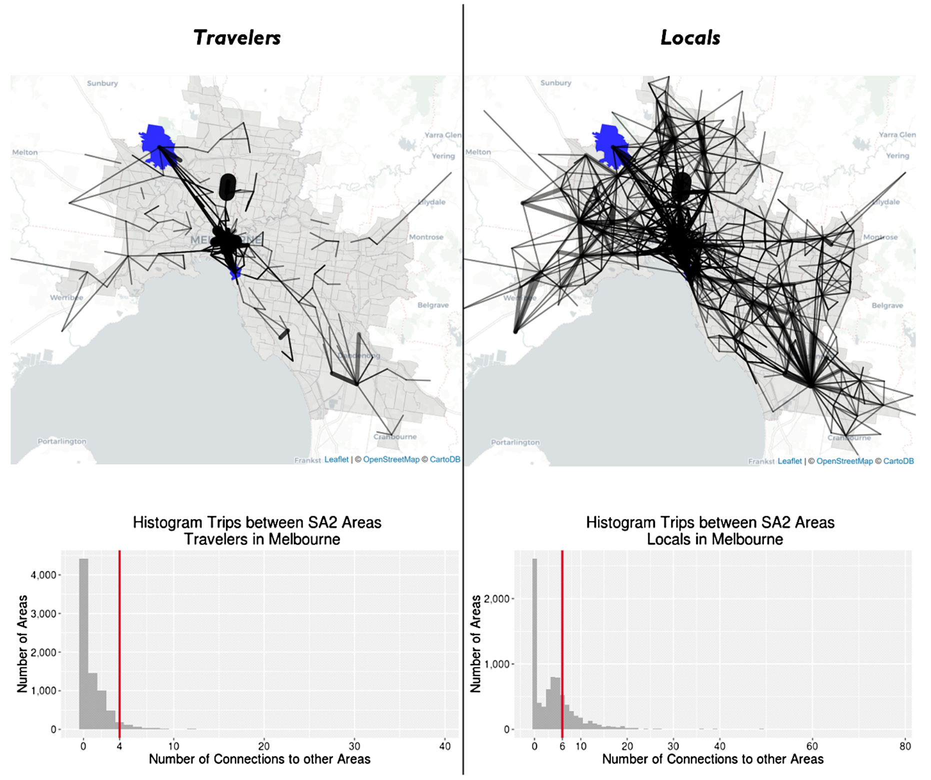

Figure 11 Top: The connectivity of SA2s in Melbourne, the line width corresponds to the number of trips between the areas. Blue color indicates the airport, the city center, and St Kilda. Bottom: Respective histograms of trips

to extract information about subgroups of a moving population in the absence of ground truth data and rigid experimental data collection protocols. We showed that location-independent behavioral features describing individuals' movement can be used to characterize distinct groups of users with distinct, and meaningful spatio-temporal movement characteristics. Using an extensive EHMD dataset tracking users in Sydney and Melbourne, we show how users with the movement characteristics of travelers and locals can be isolated with a large degree of confidence in an unlabeled dataset. We note that the features have been computed for the users in the entire dataset, while the spatio-temporal verification of the two user groups has been undertaken on a subset of the trajectories in the two cities individually. As such, the features have been totally separated from their spatial context (i.e. their absolute spatial position) and only relative position was used.

This study does not rely on additional semantic features, known to have high correlation with specific (tourist) behavior (Lew and McKercher [31]), but remaining hard to collect due to variable map data coverage. Thus, this study shows both the strengths and potential, as well as the weaknesses of using EHMD. On the one hand, mining such data has the advantage of being able to cover very large samples, or potentially even entire populations, rather than the typically small samples that are used in well-designed studies (Edwards et al. [15], Shoval and Ahas [45]). On the other hand, studies such as those reported in Edwards et al. [15] or Shoval and Ahas [45] will, apart from the tracking data, also include rich demographic data as well as qualitative data collected through interviews or questionnaires. On the one hand, as shown in the preceding section, by mining EHMD it is possible to answer questions regarding flows between tourism precincts (Kelly [28], Hayllar et al. [22]) and there is also the potential to provide answers to other important questions of tourism studies, such as those relating to typical itineraries of tourists between precincts, or potential interactions between tourists and locals (Hayllar et al. [22]). On the other hand, since EHMD has no ground truth or demographic information attached, assumptions have to be made that are plausible but might still be wrong, such as the assumption that most travelers are indeed tourists.

While the methodology can be ported to other kinds of datasets, it remains to be seen whether a model *learned* from one kind of dataset can be *transferred* to another. The evaluation of the portability of features characterizing individual clusters to other geographical contexts (countries), or datasets with different sampling rates (e.g., Call Detail Record data) needs to be undertaken. Our method may provide means to further nuance recent efforts to predict future movements of people, as recently discussed by Cuttone et al. [10].

Abbreviations

AGNES, Agglomerative Nesting; ASGS, Australian Statistical Geography Standard; CLARA, Clustering Large Applications; CV, hourly coefficient of variation; DIANA, Divisive Analysis; EHMD, Exhaust Human Movement Data; LQ, location quotient; GPS, Global Positioning System; PCA, Principal Component Analysis; SA2, Statistical Area Level 2; SL, Significant Locations.

Availability of data and materials

Data is from *Sygic* (http://www.sygic.com/gps-navigation). Due to privacy consideration regarding subjects in our dataset, we cannot make our data publicly available. The data contains human mobility at high spatio-temporal resolution from which is easy to reconstruct personal mobility patterns and infer sensitive personal information.

Competing interests

The authors declare that they do not have any competing interests.

Authors' contributions

Designed the study: LS, MT, PR. Analyzed the data: LS. Wrote the paper: LS, MT, PR, RW. All authors read and approved the final manuscript.

Author details

[1]Department of Geography, University of Zurich, Zurich, Switzerland. [2]Department of Infrastructure Engineering, The University of Melbourne, Victoria, Australia.

Endnotes

[a] *Consecutive* refers to the next day the app has been used, which does not necessarily mean the immediately following day.

[b] A move segment describes a segment of continuous movement without a stop segment. A stop segment is a series of GPS tuples with almost no movement (sum of covered distances shorter than 10 m) in a five-minute window (see also Sect. 4).

[c] We used a concave hull to compute the areal measures.

[d] See also the definition of significant locations in Sect. 5.

[e] The number of spatial clusters per user were calculated by applying a DBSCAN clustering over all SL's per user.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 25 January 2018 Accepted: 25 June 2018 Published online: 04 July 2018

References

1. Ahas R, Aasa A, Yuan Y et al (2015) Everyday space–time geographies: using mobile phone-based sensor data to monitor urban activity in Harbin, Paris, and Tallinn. Int J Geogr Inf Sci 29:2017–2039. https://doi.org/10.1080/13658816.2015.1063151
2. Asakura Y, Iryo T (2007) Analysis of tourist behaviour based on the tracking data collected using a mobile communication instrument. Transp Res, Part A, Policy Pract 41:684–690. https://doi.org/10.1016/j.tra.2006.07.003
3. Australian Bureau of Statistics (2016) Australian statistical geography standard (ASGS). In: Stat. geogr. http://www.abs.gov.au/websitedbs/d3310114.nsf/home/australian+statistical+geography+standard+(asgs). Accessed 7 Mar 2017
4. Birenboim A, Shoval N (2015) Mobility research in the age of the smartphone. Ann Assoc Am Geogr 106:283–291. https://doi.org/10.1080/00045608.2015.1100058

5. Bro R, Smilde AK (2014) Principal component analysis. Anal Methods 6:2812–2831. https://doi.org/10.1039/c3ay41907j
6. Brock G, Pihur V, Datta S, Datta S (2008) clValid: an R package for cluster validation. J Stat Softw 25:1–28. https://doi.org/10.18637/jss.v025.i04
7. Calabrese F, Diao M, Di Lorenzo G et al (2013) Understanding individual mobility patterns from urban sensing data: a mobile phone trace example. Transp Res, Part C, Emerg Technol 26:301–313. https://doi.org/10.1016/j.trc.2012.09.009
8. Calabrese F, Pereira FC, Di Lorenzo G et al (2010) The geography of taste: analyzing cell-phone mobility and social events. In: 8th international conference on pervasive computing, pervasive, pp 22–37.
9. Csáji BC, Browet A, Traag VA et al (2013) Exploring the mobility of mobile phone users. Phys A, Stat Mech Appl 392:1459–1473. https://doi.org/10.1016/j.physa.2012.11.040
10. Cuttone A, Lehmann S, González MC (2018) Understanding predictability and exploration in human mobility. EPJ Data Sci 7:2. https://doi.org/10.1140/epjds/s13688-017-0129-1
11. Datta S, Datta S (2003) Comparisons and validation of statistical clustering techniques for microarray gene expression data. Bioinformatics 19:459–466. https://doi.org/10.1093/bioinformatics/btg025
12. Dodge S, Laube P, Weibel R (2012) Movement similarity assessment using symbolic representation of trajectories. Int J Geogr Inf Sci 26:1563–1588. https://doi.org/10.1080/13658816.2011.630003
13. Dodge S, Weibel R, Forootan E (2009) Revealing the physics of movement: comparing the similarity of movement characteristics of different types of moving objects. Comput Environ Urban Syst 33:419–434. https://doi.org/10.1016/j.compenvurbsys.2009.07.008
14. Edwards D, Griffin T (2013) Understanding tourists' spatial behaviour: GPS tracking as an aid to sustainable destination management. J Sustain Tour 21:580–595. https://doi.org/10.1080/09669582.2013.776063
15. Edwards D, Griffin T, Hayllar B et al (2009) Understanding tourist 'experiences' and 'behaviour' in cities: an Australian case study. CRC for Sustainable Tourism, Gold Coast
16. George G, Haas MR, Pentland A (2014) From the editors: big data and management. Acad Manag J 57:321–326
17. Girardin F, Calabrese F, Fiore FD et al (2008) Digital footprinting: uncovering tourists with user-generated content. IEEE Pervasive Comput 7:36–43. https://doi.org/10.1109/MPRV.2008.71
18. Gong L, Liu X, Wu L, Liu Y (2016) Inferring trip purposes and uncovering travel patterns from taxi trajectory data. Cartogr Geogr Inf Sci 43:103–114. https://doi.org/10.1080/15230406.2015.1014424
19. González MC, Hidalgo CA, Barabási A-L (2008) Understanding individual human mobility patterns. Nature 453:779–782. https://doi.org/10.1038/nature06958
20. Han J, Kamber M, Pei J (2012) Data mining: concepts and techniques, 3rd edn. Elsevier, Amsterdam
21. Hayllar B, Griffin T (2005) The precinct experience: a phenomenological approach. Tour Manag 26:517–528. https://doi.org/10.1016/j.tourman.2004.03.011
22. Hayllar B, Griffin T, Edwards D (2008) Urban tourism precints: engaging with the field. In: Hayllar B, Griffin T, Edwards D (eds) City spaces-tourist places: urban tourism precincts. Butterworth-Heinemann, Oxford, pp 3–18
23. Hede AM, Hall J (2006) Leisure experiences in tourist attractions: exploring the motivations of local residents. J Hosp Tour Manag 13:10–22. https://doi.org/10.1375/jhtm.13.1.10
24. James G, Witten D, Hastie T, Tibshirani R (2013) An introduction to statistical learning. Springer, New York
25. Jiang S, Ferreira J, González MC (2015) Activity-based human mobility patterns inferred from mobile phone data: a case study of Singapore. IEEE Trans Big Data 3:208–219. https://doi.org/10.1109/TBDATA.2016.2631141
26. Kádár B, Gede M (2013) Where do tourists go? Visualizing and analysing the spatial distribution of geotagged photography. Cartogr Int J Geogr Inf Geovis 48:78–88. https://doi.org/10.3138/carto.48.2.1839
27. Kaufman L, Rousseeuw P (1990) Finding groups in data: an introduction to cluster analysis. Wiley-Interscience, New York
28. Kelly I (2008) Precincts within the urban form: relationships with the city. In: Hayllar B, Griffin T, Edwards D (eds) City spaces-tourist places: urban tourism precincts. Butterworth-Heinemann, Oxford, pp 107–126
29. Lathia N, Capra L (2011) How smart is your smartcard? Measuring travel behaviours, perceptions, and incentives. In: UbiComp'11—proceedings of the 2011 ACM conference on ubiquitous computing, pp 291–300
30. Lee J, Han J, Li X, Gonzalez H (2008) TraClass: trajectory classification using hierarchical region based and trajectory based clustering. Proc VLDB Endow 1:1081–1094
31. Lew A, McKercher B (2006) Modeling tourist movements: a local destination analysis. Ann Tour Res 33:403–423. https://doi.org/10.1016/j.annals.2005.12.002
32. Lin M, Hsu W-J (2014) Mining GPS data for mobility patterns: a survey. Pervasive Mob Comput 12:1–16. https://doi.org/10.1016/j.pmcj.2013.06.005
33. MacQueen JB (1967) Some methods for classification and analysis of multivariate observations. In: Proc 5th Berkeley symp math stat probab, vol 1, pp 281–297.
34. Mayer-Schönberger V, Cukier K (2012) Big data: a revolution that transforms how we work, live, and think
35. Neef D (2014) Digital exhaust: what everyone should know about big data, digitization and digitally driven innovation, 1st edn. Pearson FT Press, Upper Saddle River
36. Pan G, Qi G, Wu Z et al (2013) Land-use classification using taxi GPS traces. IEEE Trans Intell Transp Syst 14:113–123. https://doi.org/10.1109/TITS.2012.2209201
37. Pappalardo L, Rinzivillo S, Qu Z et al (2013) Understanding the patterns of car travel. Eur Phys J Spec Top 215:61–73. https://doi.org/10.1140/epjst/e2013-01715-5
38. Pappalardo L, Simini F, Rinzivillo S et al (2015) Returners and explorers dichotomy in human mobility. Nat Commun 6:1–8. https://doi.org/10.1038/ncomms9166
39. Parent C, Pelekis N, Theodoridis Y et al (2013) Semantic trajectories modeling and analysis. ACM Comput Surv 45:1–32. https://doi.org/10.1145/2501654.2501656
40. Phithakkitnukoon S, Horanont T, Di Lorenzo G et al (2010) Activity-aware map: identifying human daily activity pattern using mobile phone data. In: Human behavior understanding. Lecture notes in computer science, pp 14–25
41. Pihur V, Datta S, Datta S (2009) RankAggreg, an R package for weighted rank aggregation. BMC Bioinform 10:62. https://doi.org/10.1186/1471-2105-10-62

42. Reades J, Calabrese F, Ratti C (2009) Eigenplaces: analysing cities using the space—time structure of the mobile phone network. Environ Plan B, Plan Des 36:824–836. https://doi.org/10.1068/b34133t
43. Ren Y, Salim FD, Tomko M et al (2016) D-Log: a WiFi log-based differential scheme for enhanced indoor localization with single RSSI source and infrequent sampling rate. Pervasive Mob Comput 37:94–114. https://doi.org/10.1016/j.pmcj.2016.09.018
44. Rousseeuw P (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 20:53–65. https://doi.org/10.1016/0377-0427(87)90125-7
45. Shoval N, Ahas R (2016) The use of tracking technologies in tourism research: a review of the first decade. Tour Geogr 18:1–20. https://doi.org/10.1080/14616688.2016.1214977
46. Song C, Koren T, Wang P, Barabási A-L (2010) Modelling the scaling properties of human mobility. Nat Phys 6:818–823. https://doi.org/10.1038/nphys1760
47. Sygic (2016) About Sygic. http://www.sygic.com/about. Accessed 14 Nov 2016
48. Tibshirani R, Walther G, Hastie T (2001) Estimating the number of clusters in a data set via the gap statistic. J R Stat Soc, Ser B, Stat Methodol 63:411–423
49. Trasarti R, Pinelli F, Nanni M, Giannotti F (2011) Mining mobility user profiles for car pooling. In: Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining—KDD '11. ACM, New York, pp 1190–1198
50. Wang D, Pedreschi D, Song C et al (2011) Human mobility, social ties, and link prediction. In: Proc 17th ACM SIGKDD int conf knowl discov data min—KDD '11, pp 1100–1108. https://doi.org/10.1145/2020408.2020581
51. Witten IH, Frank E, Hall MA (2011) Data mining: practical machine learning tools and techniques, 3 edn. Morgan Kaufmann, San Mateo
52. Zhang D, Huang J, Li Y et al (2014) Exploring human mobility with multi-source data at extremely large metropolitan scales. In: Proceedings of the 20th ACM annual international conference on mobile computing and networking—MobiCom 2014. ACM, New York, pp 201–212
53. Zhao Z, Shaw S-L, Xu Y et al (2016) Understanding the bias of call detail records in human mobility research. Int J Geogr Inf Sci 30:1738–1762. https://doi.org/10.1080/13658816.2015.1137298
54. Zheng Y, Liu L, Wang L, Xie X (2008) Learning transportation mode from raw GPS data for geographic applications on the web. In: Proceeding of the 17th international conference on World Wide Web—WWW '08. ACM, New York, p 247
55. Zheng Y, Liu T, Wang Y et al (2014) Diagnosing New York city's noises with ubiquitous data. In: Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing—UbiComp '14 adjunct. Assoc. Comput. Mach., New York, pp 715–725