2018 IEEE Fourth International Conference on Big Data Computing Service and Applications

Macroscopic Patterns in Sparse Location Data: Identifying Mobility Prototypes

Bianca Katharina Lüders
Service-centric Networking
Telekom Innovation Laboratories
Technische Universität Berlin
Berlin, Germany
bianca.lueders@tu-berlin.de

Peter Ruppel
Service-centric Networking
Telekom Innovation Laboratories
Technische Universität Berlin
Berlin, Germany
peter.ruppel@tu-berlin.de

***Abstract*—Ubiquitous computing and location-based services are key enablers for gaining novel insights into human movement behavior on a large scale. When searching for patterns in movement behavior, several approaches have been proposed to compare location trajectories or semantic overlaps. This paper introduces the notion of *mobility prototypes*, which describe human mobility behavior at a macroscopic level and express its most significant characteristics. Based on a dataset of coarse spatiotemporal data from almost 4000 mobile users of a location-based service over a period of 14 days, we outline a comprehensive and systematic pipeline for processing and exploiting mobility traces and provide insights into the mobility behavior in the city of Berlin. We present a framework to characterize individual users precisely by employing a novel combination of discriminative features capturing the macroscopic mobility behavior, without revealing their identity or actual whereabouts, and define mobility prototypes by constructing clusters of users. The results yield valuable insights into the variety of human mobility and open up new possibilities for future mobile services.**

I. Introduction

An increasing fraction of the world population lives in cities, which requires intelligent traffic and organization systems and rises new challenges to urban planning. Accordingly, the study of urban dynamics, traffic and travel demand prediction are some of the most important research branches, and a particular important aspect is the area of *location analytics*, the science of identifying patterns among large groups of individuals and developing prediction models of human movements [1]–[7].

Urban planners need to rely on large volumes of structured high-quality data for analyses, prognoses and recommendations. Which data sources could serve as an indicator and real-time sensor for urban mobility? Many research approaches rely on taxi data [3], [4], data *actively* contributed by volunteers via surveys and manual capturing processes [8], interviews, censuses or experimental setups with dedicated hard- and software [9]. As mobile devices became an ubiquitous phenomenon, a first wave of location analytics research concentrated on data that could be gathered *passively* as a "side-product" at the operator's side of a mobile communication network. Mobile call and data transmission records, which represent the approximate location of a mobile device on the basis of the serving cell tower, are promising approaches for large-scale analyses of population-level mobility measures [5], [10], [11].

Nowadays, mobile devices are equipped with mobile data access and sensing capabilities and have advanced to full-fledged personal assistants. Novel mobile services assist the user in every situation, and especially the area of *context-aware services* attracts a lot of popularity and attention, as these provide relevant information related to the user's current situation and direct surroundings. Of all parameters which define a user's context, the *location* can be easily derived by various robust and accurate positioning technologies, which forms the basis for *location-based services* (LBS). These employ the location as a context to filter the delivered content and range from navigation and route tracking to location-based reminders and location recommender systems. While reactive LBS were only able to deliver content based on request and determine the location during an active service session, proactive LBS monitor the user's location continuously in the background. Supported by the progression of mobile operating systems, which provide interfaces for acquiring and utilizing location data in mobile applications, they supply the user with content based on predefined events [12], [13].

Thus, mobile devices equipped with GPS and location-based mobile applications do not only provide valuable content, but also capture their owner's behavior. The usage of these services allows to employ mobile devices as millions of potential sensors [1], [7], [14] and transforms physical, real-world movement into a constantly increasing asset of mobility data. Opposed to mobile call data, which are both sparse in time, since data is only generated during active usage, and in space, since the position can only be found at the granularity of cell tower positions, background tracking employed in proactive LBS is based on an explicit opt-in by its users and gathers data regularly, without additional costs and experimental setups, and potentially covers a representative sample of the population. Therefore, these accumulated traces provide a novel opportunity to examine human mobility behavior.

Following these considerations, this paper investigates the suitability of data generated by the usage of LBS to identify *macroscopic* mobility patterns in populations, with the aim to support location analytics and urban planning to obtain a

0-7695-6396-1/18/$31.00 ©2018 IEEE
DOI 10.1109/BigDataService.2018.00027

better understanding of how individuals move through large urban environments. If the data proves to be a valid indicator for macroscopic mobility, it can be employed as a large-scale, real-time indicator for urban life and answer questions such as: How many tourists are currently close to the central station? How did the share of long-distance commuters in the population change over time? Could any group be supported with additional transport connections? Additionally, as a by-product, the resulting prototypes might be employed similar to the Sinus Milieus[1], which provide target group descriptions to support marketing activities, to facilitate the development of novel, privacy-aware mobile services, that rely on an approximate user typology rather than on detailed traces, such as delivery of news, with their depth tailored to the approximate travel time, or personalized subscription tickets for local public transport.

In this paper, we present an approach to comprehend macroscopic mobility behavior based on sparse location data. For this purpose, we provide a conceptual framework to combine individual mobility features into a unified, abstract model. This allows to characterize users in a trajectory dataset by means of a novel combination of discriminative, quantitative *macroscopic mobility features*, such that a set of universal user clusters emerges. We introduce the notion of *mobility prototypes*, which represent the average behavior of these clusters, and state that they are likely to be found in any larger city and can be employed to classify users and investigate the group frequency and behavior over time. We illustrate that the features which describe these prototypes very well capture and measure the macroscopic mobility behavior without revealing the identity or whereabouts of the users. This allows to adapt services to high-level, group-wise demands. Based on a set of distance-based location updates, we present a comprehensive and systematic pipeline for processing and exploiting coarse mobility data. This pipeline includes a *preprocessing* phase for noise filtering and outlier detection, a *processing* phase for detecting stay points, deriving trips and identifying frequently visited places from sparse movement data, and a *pattern mining* phase for deriving the mobility prototypes. The city of Berlin serves as an analysis example, which illustrates several valuable, well interpretable insights into regional mobility behavior. Our main contributions are as follows:

- We specify a set of 18 features that capture the macroscopic mobility behavior of a user and form the basis for the computation of the proposed mobility prototypes (Section III). This combination is innovative and has not yet been presented.
- We present a coarse LBS-generated dataset, which captures the mobility of more than 4000 users, and outline a systematic processing methodology (Section IV).
- We illustrate the main findings of our empirical study, provide insights into population-level mobility behavior and characterize the resulting mobility prototypes (Section V).

[1]https://www.sinus-institut.de/en/sinus-solutions/sinus-milieus/

II. Related Work

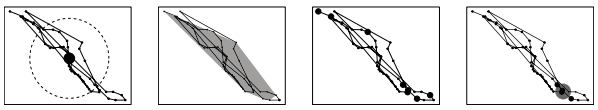
Besides well-known works concerning human mobility patterns [15]–[17], a few recent survey papers might serve as a broader entry point for research in trajectory data mining [7], [14], [18]–[20]. Many approaches focus on the extraction of discriminative features for the purpose of transport mode detection [21], the detection of mental health issues [6], or behavior clusters according to weekday and weekend activities [8]. Cheng et al. [22] present a highly insightful analysis of check-in data and demonstrate periodic behaviors and the influence of geographic, economic and social aspects on mobility patterns. Ashbrook and Starner [9] utilize continuously recorded GPS data, detect stays based on time gaps, identify relevant locations with a variant of k-Means clustering and incorporate these into a predictive model. Based on anonymized cellular network data, Isaacman et al. [5] demonstrate an accurate and efficient method to identify semantically meaningful locations using clustering and logistic regression methods, validate the results based on ground truth provided by volunteers, calculate commute distances and estimate the corresponding carbon footprints. With a focus on traffic analysis, Bischoff et al. [3] use taxi trajectories to analyze travel behavior and vehicle supply of the Berlin taxi market, and Zheng et al. [4] detect flawed urban planning using the GPS trajectories of taxicabs traveling in urban areas in Beijing. To our knowledge, no group has yet presented a pipeline for handling coarse position data from LBS, which requires dedicated approaches for stay point and personal location identification. Similarly, no approach could be found to derive macroscopic mobility features from coarse movement data and employ these as a basis to derive mobility prototypes.

III. Macroscopic Mobility Features

We introduce a new combination of features designed to capture a user's individual macroscopic mobility behavior. These features allow to aggregate a large set of mobility data in a *user description matrix* (UDM), in which each row describes one of n users in d dimensions, and include both common measures of human mobility as well as newly defined statistics to summarize macroscopic mobility behavior. The features are designed to be applicable to any research setting, non-dependent on the underlying geographical region, address the spatial nature of human mobility and measure clearly defined aspects. Thus, the set combines information about a user's *basic trajectory*, his *stay points*, *trips* and *personal locations*, his overall *mobility*, and his *temporal behavior* (Table I and Figure 1). While the stay points are those positions where the user remained for a certain period, such that trips reflect the movement between two stops, the personal locations represent a user's frequently visited places and may disclose information about home and work locations.

In the first category, the commonly employed *radius of gyration* r_g is incorporated, which is a unidimensional measure to quantify the typical coverage area of a user's trajectory and the extent of moving away from the *center of mass* c_m [10], [22], [23], the Euclidean mean of his positions. To

avoid influence of the recording strategy, r_g is computed on the basis of the stay points. González et al. demonstrated that r_g increases with time and reaches a saturation [16]. Figure 1(a) shows a sample plot of a user's recordings, the c_m and r_g centered around it. To determine the underlying *mobility model*, a fit to the distribution of spatial distances between stops can be revealing. Many authors found this to be approximated by a power law, indicating that the user follows a *Lévy Flight*, drawing the *displacements* from a heavy-tailed distribution [15], [22]. Thus, the power law fit coefficient α and x_{min}, which indicates the lower bound of the distribution for the power law behavior [24], for both the distributions of displacements and waiting times are incorporated in the set. As much as α_{disp} indicates how slowly the distribution decays, i.e. how pronounced the heavy tail is, α_{wt} gives insights into how much long periods of rest attenuate the user's spatial spread. Similarly, the *return probability* is a part of the feature set, which measures periodic behavior and was found to be characterized by peaks at daily and weekly intervals, which can be considered to be a factor for a stabilization of r_g [16]. To estimate a user's probability to return to his personally preferred locations, the times between visits were divided by 24 hours, the remainders were calculated and the percentage of remainders below 3 and within a time frame of $19 - 29$ hours computed. Besides these common measures, the feature set also captures how often and for how long a user visited his personal locations. The overall travel behavior is integrated as well, with the aim to differentiate between users who tend to travel at a constant speed and those who are likely to change their means of transport, captured by the median average deviation (*MAD*) of the trip speed. Additionally, the concept includes a *touristic score*, which takes into consideration the regional points of interest and the amount of times a given user had a stay point in the proximity of a touristic location. The resulting values are normalized by the highest observed score ($s_{tourist} \in [0, 1]$) such that the resulting values are comparable across geographical regions. Finally, to differentiate between visitors and residents, the set contains a *city time share*, for which the hours spent within the city are aggregated.



a) b) c) d)

Fig. 1: Illustrations of the macroscopic mobility feature bases. (a) Sample trajectory, c_m and r_g (b) Covered area (c) Stay points (d) Location as cluster of stay points.

IV. The Trajectory Processing Pipeline

This section presents the systematic pipeline of our framework, illustrated in Figure 2, starting with a dataset and proceeding with a preprocessing phase, i.e., quality and plausibility checks for filtering noise and removing test users, and an outlier detection. Following, the processing phase extracts the individual stay points, which are the basis to segment the trajectories into trips. These allow to derive information about frequently covered distances. Clustering individual stay points into locations then allows to determine personal, frequently visited locations. The result of the pipeline, the UDM, is the basis for clustering macroscopically similar users into groups and shape the resulting mobility prototypes.

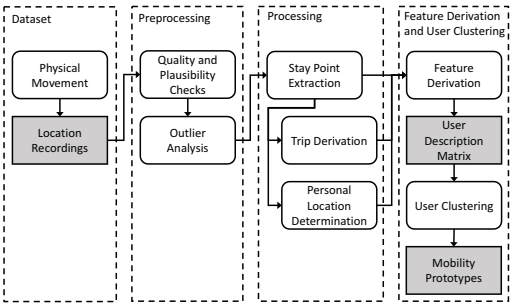


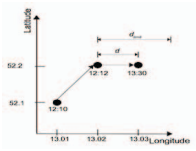
Fig. 2: Stages of the trajectory processing pipeline

A. Dataset

The dataset was provided by the operators of an opt-in based, proactive location-based service for mobile devices. To supply its users with relevant, interesting, up-to-date and location-related information, the application needs to query the backend server whenever the device has covered a certain distance, which comprises the transmission of the current position. Each of these resulting *location updates* describes a user position recording by means of a timestamp, a user pseudonym and the geographical location.

B. Preprocessing

A set of quality and plausibility checks is employed, including the adjustment of sudden jumps which can occur due to switching positioning systems. Due to temporally very different daily routines of users, the outlier detection focuses on the spatial spread. Taking into account the structure and generation mechanisms of position data, common approaches are combined into a two-step outlier detection procedure, consisting of a parametric removal of users with the highest deviation from the overall center of mass based on the Mahalanobis distance of their recordings and a distance-based non-parametric detection to detect outlying recordings within individual traces.



Distance	Assumed Speed
< 3 km	5 km/h
3 - 20 km	25 km/h
20 - 50 km	50 km/h
> 50 km	130 km/h

a) b)

Fig. 3: Main concepts. (a) Stay point detection (b) Assumed mean speed depending on spatial distance.

TABLE I: Macroscopic mobility features

Basic Trajectory	Stay Points & Trips	Personal Locations	Mobility	Temporal Behavior
• Total covered distance • Total covered area as captured by the outermost positions	• Number of trips • Median speed • Median duration • Median distance • MAD speed • Touristic Score	• Number of locations • Average distance between locations • Average time spent at the top locations	• Radius of Gyration • Displacements power law fit estimates • Waiting times power law fit estimates • Return probability	• City time share

C. Processing

1) Stay Point Extraction: To turn the trajectories into sequences of meaningful places, a first step is to identify *stay points*. The general idea of stay point extraction is that trajectories can be recorded by positioning systems based on several different underlying *recording strategies*, including *time*, *change*, *location* and *event-based* recording. It is common to segment trajectories further by defining them as a sequence of *stops* and *moves*, i.e. time sub-intervals where the object stays fixed at a certain point in space or changes its position [25]. Trajectories captured with an event-based recording strategy mostly conform with this idea. In contrast, when the data was recorded in a time or change-based manner, many positions were recorded solely during the movement, and it is necessary to extract meaningful stops and define moves as the proceeding among them. Li et al. [26] first proposed to identify these points based on temporal and spatial distance thresholds, which is adapted here to take the present distance-based recording strategy into account. Hence, if two positions of a user are apart by more than a given time interval t_{limit}, but at the same time closer to each other than a given spatial limit d_{limit}, the location of the first is registered as the position of the stay point and its timestamp is defined to be the arrival time. To compute the departure time, the approach subtracts the approximated travel time between the positions, estimated based on their spatial distance and related mean speeds (Figure 3(b)), from the arrival time at the second position (Figure 3(a)). The observed spatial distance d is smaller than d_{limit} and the time between the two upper recordings is larger than a temporal limit $t_{limit} = 30min$. Assuming that the user moves at a given speed that indicates that the distance from the central to the right point can be covered in 10 minutes, this would yield the following first stay point for user u: $sp_u^1 = \{userID = u, longitude = 13.02, latitude = 52.2, arrival = 12:12, departure = 13:20\}$. Based on related approaches, each user's stay points are extracted based on a temporal limit of 10 minutes and a spatial limit of 1.5 kilometers.

2) Trip Derivation: The stay point extraction concept allows to define trips as the movement between two consecutive stay points of a user, represented by the user ID, the origin, the start time, the destination and the arrival time, the beeline distance, the travel time and the weekday of travel. Since the framework focuses on macroscopic movement patterns, it suffices to model trips as straight lines between stay points.

3) Personal Location Determination: Next, a user's stay points are clustered into his frequently visited locations, which supposedly represent meaningful places such as home and work. Due to inaccuracies in GPS positioning, positions that in fact represent visits to the same physical location may differ in their recorded geographic coordinates, and it is necessary to employ a single coordinate pair as a representative [6]. Similar approaches are Hartigan's leader algorithm [5] and variants of k-Means [9]. At this point, the goal is to employ a clustering algorithm which results in compact clusters, allows for arbitrary shapes, is able to identify points as noise and accepts a reasonable notion of distance. Hence, DBSCAN [27], which does not require to define the number of clusters a priori, is appropriate. It generates a non-hierarchical partitioning of the data, based on the key idea that each point p of a cluster should be surrounded by a neighborhood $N_\epsilon(p)$ of a given radius ϵ that contains at least a minimum number $MinPts$ of points. Following the suggestions by Ester et al., the minimum cluster size is set to 3 and ϵ to 0.008, which corresponds to distances between 500-750m.

D. Feature Derivation and User Clustering

Given that the measures are on very different scales, they are standardized using *z-score* scaling. Since the clusters should be compact and convex, without a definition of noise, and with a flat partitioning, the approach employs k-Means. This algorithm initializes a cluster center iteration and generates Voronoi-diagram shaped clusters. Since the dimensions are scaled and not comparable in their real-world meaning, the Euclidean distance is used. To approach common issues with k-Means, we test for outliers using a parametric outlier detection and perform several iterations to verify the stability of the results. Since it is necessary to decide about the number of clusters a priori, we employ an Elbow Plot, which plots the number of clusters k against the total within-cluster sum of squares, Dunn's index [28] and the average Silhouette index [29]. Finally, we perform the clustering based on the scaled user description matrix, attach the cluster association to the original matrix and define the cluster centers as the resulting mobility prototypes.

V. Empirical Results

A. Dataset

The preprocessing phase identified 20.30% of the users of the original data set as invalid, such that the data collection

resulted in location data of 3985 users, covering two consecutive weeks from Monday, May 23, to Sunday, June 5 of 2016. The locations were recorded with a distance-based recording strategy, with a resolution of one kilometer on average. The users were selected by requiring them to have at least one location inside the city of Berlin. Table II offers some general data characteristics.

TABLE II: General data characteristics

#users	3,985
#positions	949,887
positions per user	on average 238.40, median 135.00
positions per user per day	on average 20.33, median 9.00
longitude range	[-16.55, 55.36], median 13.30, $\sigma = 1.95$
latitude range	[24.67, 60.51], median 52.47, $\sigma = 1.29$
distance between positions	on average 3.075 km, median 0.683 km
time between positions	on average 1.28 h, median 0.03 h
speed between positions	on average 62.20 km/h, median 43.41 km/h

The distribution of positions per user is highly right-skewed, with a minimum of 6 and a maximum of 4464 recordings. Both the distribution of beeline distance and the time between subsequent position recordings are right-skewed and coined by outliers. While 94.86% of the positions are less than 10km apart and 60.86% less than 1km, the maximum distance between two events is more than 4000km. Similarly, 85.36% of the subsequent positions are less than half an hour apart. Figure 4(a) illustrates the distributions of longitude and latitude, where the vertical lines indicate the bounding box of Berlin. Both distributions appear nearly normal and more than 83% of the positions are less than one standard deviation away from the respective mean. The heatmap (Figure 4(b)) indicates that a large part of the positions was generated in the city center and along the urban highway. Figure 5(a) visualizes the amount of positions per hour and uncovers both the general regularity of the recordings and the aggregated daily patterns. Monday mornings and Friday evenings are extremely busy time periods, connected by a midweek activity attenuation, which might be related to commuters who either arrive in Berlin or leave the city during the week. Furthermore, the activity patterns of the weekends are different from those during weekdays and the activity is generally higher during daytime than at night. While the weekdays repetitively exhibit two significant daily peaks, the weekend activity starts much later and shows a continuous plateau from 10am to 6pm (Figure 4(c)).

B. Data Processing

1) Stay Points: The proposed algorithm extracts 69,442 stay points, of which 72.78% are in Berlin. On average, each user has 17.45 stay points, with a maximum of 83. The pause times spread around a mean value of 5.51h, which can be associated with work hours, while 50% of the stays are shorter than one hour, potentially corresponding to shopping stops, coffee breaks and waiting times for public transport.

2) Trips: The stay points allow to derive a set of 65,422 trips, with an average of 16.7 trips per user, of which 66.85% start and end in and 21.31% start and end outside of Berlin. Figure 6 provides a Chord Diagram of trips within the city. The edges going out of and coming into Charlottenburg-Wilmersdorf are highlighted. This illustrates how few trips go directly from the most western to the eastern districts and indicates the large portion of trips remaining within each district, which may be suggestive of a high bond of the city's inhabitants to their preferred areas.

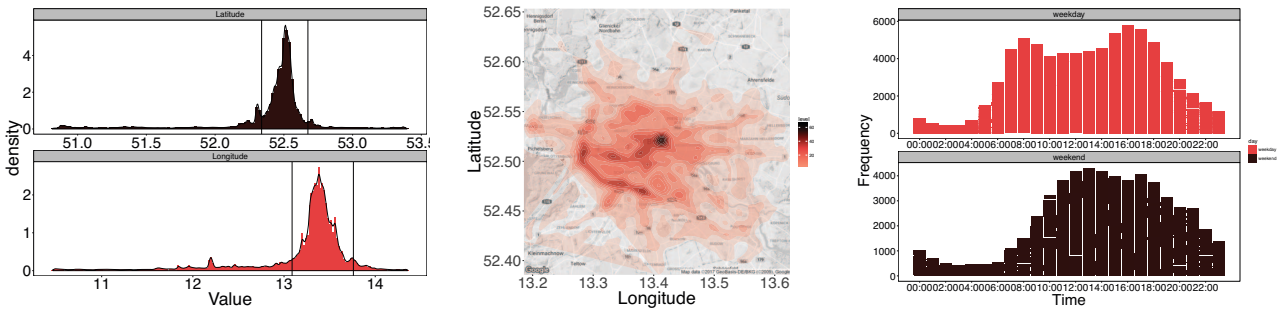
The trips within Berlin have an average distance of 4.7km, with a maximum of 35km, which covers the entire city. Trips which start and end within the city take one hour on average. Considering the entire set of trips, these values increase significantly, such that the mean distance raises to 20km and the mean speed to 7km/h. Differentiating the trips between starting on a weekday or the weekend, we find not only more trips on weekdays, but also find them to have both a longer duration and covering a longer distance than weekend trips, as summarized in Table III, where the medians of distance and duration are given. The results strongly reinforce the idea of many long-distance commuting trips during the week.

TABLE III: Weekday trip details

Day	**#active users**	**#trips**	**distance [km]**	**duration [h]**
Sunday	2322	3.36	0.82	5443
Monday	3282	4.03	1.38	11181
Tuesday	3116	3.94	2.19	9283
Wednesday	3250	3.80	1.68	10148
Thursday	3267	3.83	1.73	10542
Friday	3322	3.95	1.46	11067
Saturday	2774	3.59	1.33	7758

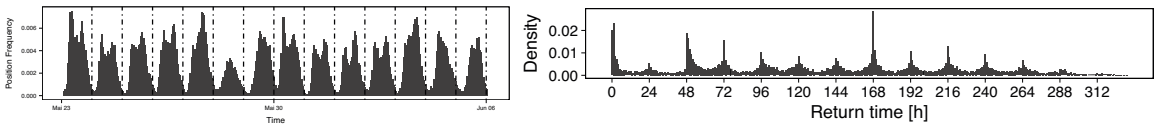
3) Personal Locations: The framework succeeded in finding 5,463 locations for 77.94% of the users, of which 51.06% have exactly one location. On average, each location is visited four times by the corresponding user, with a visiting time of 6-8 hours. Of the locations, 77.02% are in Berlin, and 2215 users have locations only within the city, while 313 users have personal locations as well inside as outside of the city. Similar to other studies, this implicates that even though the users visit many different places, they also show a focus on a small set of frequently visited ones [5]. Figure 5(b) visualizes a histogram of the return times to previously visited locations and exhibits high peaks at multiples of 24 hours, indicating a strong daily return probability. A peak at 168 hours hints at an especially strong weekly return probability.

4) Macroscopic Mobility Features: Finally, the approach derives the mobility features for users with at least one personal location. The resulting univariate feature distributions, of which most are very right-skewed, are shown in Figure 7, along with their median and mean values. Some ranges have been shortened for easier plotting. The resulting r_g show a highly right-skewed distribution, 60.95% of the users have a r_g below 10km. In contrast to the median (7.23km), the average value is very high (45.46km), indicating that, while the activity of most users is confined to a limited neighborhood, some users regularly cover large distances. An outlier detection of the scaled UDM based on the Euclidean distance to the average user discards four extremely outlying users who could



a) b) c)

Fig. 4: Spatial and temporal data analysis. (a) Distributions of latitude and longitude (b) Heat map of positions in Berlin (c) Average daily pattern of positions.



a) b)

Fig. 5: Temporal data analysis. (a) Positions per Hour (b) Return times, rounded to full hours.

Fig. 6: Chord diagram of trips in Berlin

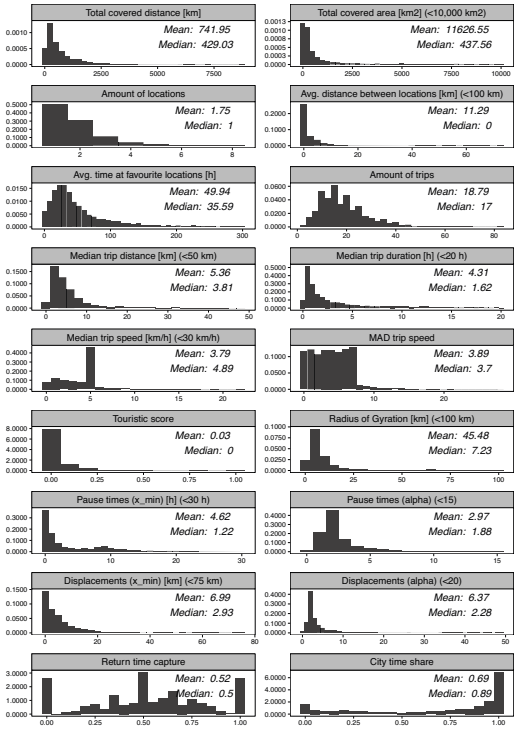


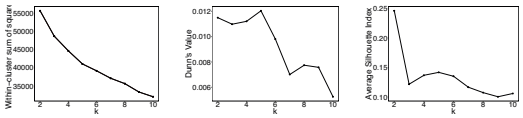
Fig. 7: Univariate feature distributions

bias the clustering. An analysis of linear correlations among the features shows that the total covered area, the total covered distance and r_g are positively related.

C. Mobility Prototypes

The results of the methods described in Section IV-D are shown in Figure 8. The Elbow Plot exhibits a bend at $k = 5$, where both Dunn's and the Silhouette index show a peak as well. This indicates that five clusters are appropriate for this dataset. Applying k-Means with $k = 5$ to the scaled UDM and computing the cluster centers yields the mobility prototypes illustrated in the radar chart in Figure 9, where negative values point towards the center. A description of the most telling cluster characteristics and the percentages of users assigned to each group are provided below.

1) Yellow cluster, 7.58%, Long-distance frequent travelers: This cluster is the smallest one and has the highest mean



a) b) c)

Fig. 8: Index criteria for k. (a) Elbow Plot (b) Dunn's Index (c) Silhouette Index.

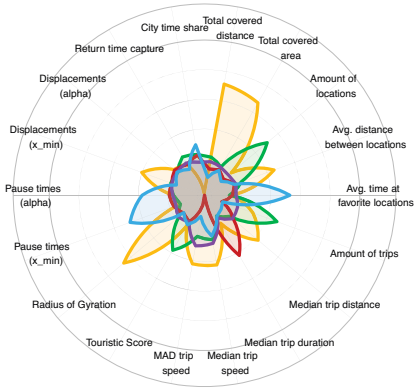


Fig. 9: Mobility prototype shapes along 18 dimensions

covered distance (2933.68km), corresponding to the highest total covered area (98145.40km^2). The users, who have 1.69 personal locations on average, also exhibit the highest distance between those, of more than 86km, the largest trip distances and speeds, and the highest r_g (278.39km), combined with the lowest city time share. These users are not only likely to be long-distance, but also frequent travelers, such as consultants with a home in Berlin.

2) Green cluster, 16.06%, Within-city commuters: These users show the highest amount of locations (3.27), the highest amount of trips (32.03), the highest touristic score (0.10), the highest return time capture (0.60) and a normal city time share of 0.75, combined with a comparatively low radius of gyration. The amount of trips, related to the amount of locations, could indicate that these users mostly move between a few close locations, of which one could be in central Berlin, explaining the touristic score, and the other more in the residential areas.

3) Red cluster, 24.77%, Neighborhood-focused residents: This cluster stands out due to the smallest number of trips (12) and a very high median trip duration of nearly 12 hours, combined with the lowest trip speed and its lowest variation, the second-smallest amount of locations as well as the second-smallest average distance between locations, the lowest touristic score and a high city time share. This is suggestive of a residential cluster with hardly any movement, but which is different from the purely home-focused residents due to longer trips.

4) Violet cluster, 37.54%, Regular-distance commuters: This cluster, the largest one, appears to be a more location-based version of the yellow cluster, with comparably high median trip speeds and a high radius of gyration. Also, this cluster is similar to the red cluster, but exceeds it in the total covered distance and area and the amount of trips. These users might cover regular commute distances on a smaller scale, potentially across the city or to a home location not far away from the city.

5) Blue cluster, 14.03%, Home-focused residents: This cluster stands out due to the largest city time share (0.87), the outstandingly high average time at favorite locations of more than 125 hours, the smallest covered distance (205km) and the smallest r_g (12.60km). This implies a movement in a very confined region, combined with few trips, low median trip distances and durations and the smallest average number of locations. These users are likely to be residents of Berlin, with short daily routes and a high bond to the home area.

VI. Discussion

We pinpoint possible limitations and biases in order to promote a careful attitude towards the results. First, some *limitations* root in the varying accuracy of the recorded positions, which highly depend on the availability of GPS signals. Due to the recording strategy, the data is comparably sparse, both in the temporal and the spatial dimension, and not based on uniform sampling rates as in comparable studies [3], [9]. Since it is not possible to backtrack with meter accuracy the specific location visited at any time, a microscopic mobility analysis is unfeasible; since no demographic information about the users of the mobile application is available, it is not possible to ascertain whether the data is a representative sample. Additionally, the dataset is subject to a few *biases* - it includes only users who own a GPS-enabled mobile device and have one of the applications installed that integrate the LBS. Non-users might reveal different mobility characteristics. Since LBS are mostly integrated in on-the-go services, the dataset might be biased towards highly mobile users. Furthermore, the data covers only a limited time span, a larger or different recording period could yield different results.

VII. Conclusion and Future Work

While early challenges in the area of mobility research lied in gathering large-scale, comprehensive and accurate data, advances in location acquisition, mobile computing technologies and the increasing usage of location-aware devices generated massive amounts of spatial movement data [18], [20]. The presented results demonstrate that LBS represent an promising way to gather large-scale mobility data. Building on a theoretical foundation and accompanied by a variety of insights, we present a framework to utilize data from LBS to extract underlying mobility prototypes, which reflect the mobility behavior of a part of the population. The framework includes a comprehensive data processing pipeline, covering quality and plausibility checks and outlier analysis, concepts to extract stay points, derive trips and identify personal locations, and a set of features which capture the individual macroscopic mobility behavior in a solid and precise way.

The presented feature concept is designed to be applicable to comparable datasets, and each feature describes a slightly different aspect, which is relevant both isolated as well as in a combined fashion. Based on a sample of users in Berlin, our methodology identifies five clearly separated clusters with specific movement characteristics. Since both the group sizes and shapes remain stable over several clustering iterations, we suppose that they are inherently present in the data. Even though the lack of an associated ground truth for behavior-based user groups currently prevents evidence, we hypothesize that the observed patterns occur with a significant cumulation, such that we can conclude that LBS-generated data are a valid data source and serve as an indicator for urban mobility. Additionally, the characterization of individuals with summary features can serve as a basis for the development of novel, privacy-enhancing mobile services and applications, which adjust to the macroscopic mobility prototype of a user.

The presented approach offers several possibilities for future work, of which the first is a sound evaluation of the results. In the context of an internal evaluation, the entire dataset could be employed as ground truth, such that it can be verified whether a sample would yield comparable results. To assess the stability of the algorithm and to generalize the methodology and the results, we could compare the results for two structurally similar databases, related to comparable cities, or two entirely different geographical backgrounds. On the other hand, an external evaluation would require verified knowledge about the daily mobility behavior of the users under consideration, which might be obtained via surveys, detailed GPS traces and publicly available demographic data. To our knowledge, no approach could be found which could serve as a basis for a comparison of user clusters and their frequency. To realize application scenarios which replace the precise location recordings with summaries, architecture, data storage and processing solutions need to be evaluated. This paper has not yet exploited all of the mobility knowledge that can be gathered from the dataset, and future work could consider a more extensive exploratory data analysis, additional measures for correlation, test for non-linear relationships between the features, employ more robust clustering algorithms and also raise the question whether users change between mobility prototypes - for example, we suggest to investigate a two-dimensional mobility prototype definition, which differentiates between weekday and weekend behavior.

REFERENCES

[1] I. Leontiadis, R. Stanojevic, A. Lima, D. Wetherall, H. Kwak, and K. Papagiannaki, "From Cells to Streets: Estimating Mobile Paths with Cellular-Side Data," in *CoNEXT'14*. Sydney: ACM, 2014.

[2] F. Girardin, J. Blat, F. Calabrese, F. Dal Fiore, and C. Ratti, "Digital footprinting: Uncovering tourists with user-generated content," *IEEE Pervasive Computing*, vol. 7, no. 4, pp. 36–44, 2008.

[3] J. Bischoff, M. Maciejewski, and A. Sohr, "Analysis of Berlin's taxi services by exploring GPS traces," *2015 International Conference on Models and Technologies for Intelligent Transportation Systems, MT-ITS 2015*, no. December 2012, pp. 209–215, 2015.

[4] Y. Zheng, Y. Liu, J. Yuan, and X. Xie, "Urban Computing with Taxicabs," *Proc. 13th international conference on Ubiquitous computing - UbiComp '11*, pp. 89–98, 2011.

[5] S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, M. Martonosi, J. Rowland, and A. Varshavsky, "Identifying Important Places in People's Lives from Cellular Network Data," *Pervasive Computing*, vol. 6696, pp. 133–151, 2011.

[6] L. Canzian and M. Musolesi, "Trajectories of Depression : Unobtrusive Monitoring of Depressive States by means of Smartphone Mobility Traces Analysis," *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing.*, pp. 1293–1304, 2015.

[7] D. Naboulsi, M. Fiore, S. Ribot, and R. Stanica, "Large-scale mobile traffic analysis: A survey," pp. 124–161, 2016.

[8] S. Jiang, J. Ferreira, and M. C. González, "Clustering daily patterns of human activities in the city," in *Data Mining and Knowledge Discovery*, vol. 25, no. 3, 2012, pp. 478–510.

[9] D. Ashbrook and T. Starner, "Using GPS to learn significant locations and predict movement across multiple users," *Personal and Ubiquitous Computing*, 2003.

[10] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of predictability in human mobility." *Science (New York, N.Y.)*, vol. 327, pp. 1018–1021, 2010.

[11] N. E. Williams, T. A. Thomas, M. Dunbar, N. Eagle, and A. Dobra, "Measures of human mobility using mobile phone records enhanced with GIS data," *PLoS ONE*, vol. 10, no. 7, 2015.

[12] S. Rodriguez Garzon and B. Deva, "On the Evaluation of Proactive Location-Based Services," in *Computer Software and Applications Conference (COMPSAC)*, vol. 2, 2015, pp. 585–594.

[13] A. Küpper, G. Treu, and C. Linnhoff-Popien, "TraX: A device-centric middleware framework for location-based services," *IEEE Communications Magazine*, vol. 44, no. 9, pp. 114–120, 2006.

[14] V. D. Blondel, A. Decuyper, and G. Krings, "A survey of results on mobile phone datasets analysis," pp. 1–55, 2015.

[15] D. Brockmann, L. Hufnagel, and T. Geisel, "The scaling laws of human travel." *Nature*, vol. 439, no. 7075, pp. 462–5, 2006.

[16] M. C. González, C. A. Hidalgo, and A.-L. Barabási, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, 2008.

[17] A.-L. Barabasi, "The Origins of Bursts and Heavy Tails in Human Dynamics," *Nature*, vol. 435, pp. 207–211, 2005.

[18] Y. U. Zheng, "Trajectory Data Mining : An Overview," *ACM Trans. Intell. Syst. Technol.*, vol. 6, no. 3, pp. 1–41, 2015.

[19] Z. Feng and Y. Zhu, "A Survey on Trajectory Data Mining: Techniques and Applications," *IEEE Access*, vol. 4, pp. 2056–2067, 2016.

[20] J. Mazimpaka and S. Timpf, "Trajectory data mining-A review of methods and applications," *Journal of Spatial Information Science*, pp. 1–45, 2016.

[21] Y. Zheng, Y. Chen, Q. Li, X. Xie, and W.-Y. Ma, "Understanding transportation modes based on GPS data for web applications," *ACM Transactions on the Web*, vol. 4, no. 1, pp. 1–36, 2010.

[22] Z. Cheng, J. Caverlee, K. Lee, and D. Z. Sui, "Exploring Millions of Footprints in Location Sharing Services," *Icwsm*, vol. 2010, no. Cholera, pp. 81–88, 2011.

[23] P. Ranacher and K. Tzavella, "How to compare movement? A review of physical movement similarity measures in geographic information science and beyond," *Cartography and Geographic Information Science*, vol. 41, no. November 2015, pp. 286–307, 2014.

[24] A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-Law Distributions in Empirical Data," *SIAM Review*, vol. 51, no. 4, p. 661, 2009.

[25] S. Spaccapietra, C. Parent, M. L. Damiani, J. A. de Macedo, F. Porto, and C. Vangenot, "A conceptual view on trajectories," *Data and Knowledge Engineering*, vol. 65, no. 1, pp. 126–146, 2008.

[26] Q. Li, Y. Zheng, X. Xie, Y. Chen, W. Liu, and W. Ma, "Mining user similarity based on location history," *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems*, no. c, p. 34, 2008.

[27] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," *Second International Conference on Knowledge Discovery and Data Mining*, pp. 226–231, 1996.

[28] J. C. Dunn, "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters," *Journal of Cybernetics*, vol. 3, no. 3, pp. 32–57, 1973.

[29] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, no. C, pp. 53–65, 1987.