City Research Online

City St George's, University of London

Citation: Andrienko, N., Andrienko, G., Fuchs, G. & Jankowski, P. (2016). Scalable and privacy-respectful interactive discovery of place semantics from human mobility traces. INFORMATION VISUALIZATION, 15(2), pp. 117-153. doi: 10.1177/1473871615581216

This is the accepted version of the paper.

This version of the publication may differ from the final published version. To cite this item please consult the publisher's version.

Permanent repository link: https://openaccess.city.ac.uk/id/eprint/14193/

Link to published version: https://doi.org/10.1177/1473871615581216

Copyright and Reuse: Copyright and Moral Rights remain with the author(s) and/or copyright holders. Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge, unless otherwise indicated, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way. For full details of reuse please refer to City Research Online policy.

City Research Online: http://openaccess.city.ac.uk/ publications@citystgeorges.ac.uk

Scalable and Privacy-respectful Interactive Discovery of Place Semantics from Human Mobility Traces

Natalia Andrienko[1,2], Gennady Andrienko[1,2], Georg Fuchs[1], and Piotr Jankowski[3,4]

[1] Fraunhofer Institute IAIS, Sankt Augustin, Germany

[2] City University London, UK

[3] San Diego State University, USA

[4] Institute of Geoecology and Geoinformation, Adam Mickiewicz University, Poznan, Poland

Abstract

Mobility diaries of a large number of people are needed for assessing transportation infrastructure and for spatial development planning. Acquisition of personal mobility diaries through population surveys is a costly and error-prone endeavour. We examine an alternative approach to obtaining similar information from episodic digital traces of people's presence in various locations, which appear when people use their mobile devices for making phone calls, accessing the internet, or posting georeferenced contents (texts, photos, or videos) in social media. Having episodic traces of a person over a long time period, it is possible to detect significant (repeatedly visited) personal places and identify them as home, work, or place of social activities based on temporal patterns of a person's presence in these places. Such analysis, however, can lead to compromising personal privacy. We have investigated the feasibility of deriving place meanings and reconstructing personal mobility diaries while preserving the privacy of individuals whose data are analysed. We have devised a visual analytics approach and a set of supporting tools making such privacy-preserving analysis possible. The approach was tested in two case studies with publicly available data: simulated tracks from the VAST Challenge 2014 and real traces built from georeferenced Twitter posts.

INTRODUCTION

Information about human daily mobility, i.e., where people travel, when, and why, is necessary for transportation management, urban planning, and public health studies. This kind of information can be obtained from daily mobility diaries of a large sample of people. The way of using mobility diaries is briefly described below.

The use of daily mobility diaries

In planning any changes in transportation infrastructure, public transportation, or land use in populated areas, it is necessary to obtain realistic predictions of how the changes may affect the daily lives of the population, in particular, their daily mobility behaviors, and how well the changed environment will suit the daily needs of the people. For this purpose, it is necessary to estimate how many people at different times of the day need to get to the places of their work or study, to schools or kindergartens of their children, to places for shopping, health care, services, entertainment and recreation, and back to their home places. This is done in the following way.

A large set of personal mobility diaries from a sample of the population is obtained through a survey campaign. In the diaries, the responders describe at what times of the day they typically go to what kinds of places and for what purposes. A planner generates a so called synthetic population where each virtual individual represents one person from the real population currently living or expected to live in the area to be affected by the planned changes. The collected diaries are randomly distributed among the virtual individuals and treated as their personal plans for a day. The place types referred to in the diaries are substituted by concrete locations in the area based on the current or planned land use distribution. Then agent-based simulation methods [1] are used to simulate movements of the virtual individuals through the planned environment for implementing their personal plans. The individual movements are aggregated into collective flows. The simulation may uncover bottlenecks in the transportation infrastructure, show where people's time may be lost, reveal unused facilities, etc. Based on such findings, the planner can understand what needs to be improved and in what way. Similar studies are undertaken to forecast the impacts of possible changes in the population number or structure.

The data acquisition problem

Personal mobility diaries are traditionally acquired through population surveys, which is a costly and error-prone endeavour. Therefore, researchers have been seeking alternative ways to obtain similar information, preferably, from already existing data containing traces of people's presence and activities in different places at different times, such as mobile phone use records. Given that mobile phones are actively used throughout a day, the daily mobility can be reasonably well reflected in the data. Since the mobile phone data are not widely accessible, other publicly available data with similar properties, in particular, georeferenced posts in Twitter and other social media, have also been investigated. Such posts are most often sent from mobile devices; hence, their coordinates and time stamps can be used to trace the users' mobility.

There is a large content gap between daily mobility diaries and automatically collected data with time-stamped positions of people. The latter lack semantics, which makes it difficult to ascertain whether a phone call or Twitter post has been made from home, work place, public transport, grocery store, or other type of location. Hence, to be able to reconstruct diary-like information from mobility data, it is necessary to extract significant locations and determine their person-specific meaning or purpose, such as home, work, shopping, sports, social activities, etc.

An earlier work [2] demonstrated a possibility of extracting frequently visited places from raw position data of a person (specifically, GPS tracks) covering a long time period, and identifying the meanings of these places by analysing temporal patterns of the person's presence. However, what is easy to do with data from a few people becomes very problematic for hundreds or thousands of people. Moreover, analysing such data at the individual level can reveal sensitive personal information, thus compromising personal privacy. Hence, the challenge is to find approaches to deriving place semantics for a large number of people by analysing entire population and groups (overall and intermediate levels [3]) instead of individuals.

A further difficulty is posed by the episodic character of mobile phone use records, georeferenced tweets, and similar data [4][5]. Persons' positions are recorded only when specific events occur: starting a call, sending an SMS, accessing a web page, or sending a tweet. For the times between the events, the whereabouts of the people are unknown. This means, in particular, that the time stamps of entering and leaving places and the duration of staying are not known and cannot be used in the analysis, which is different from analysing GPS tracks with fine temporal

resolution [2]. With such episodic mobility data, the place semantics need to be derived from temporal distributions of events recorded at those places. The feasibility of this concept has been demonstrated on the example of a few individuals considered separately [6]. The challenge is to extend it to many individuals analysed simultaneously.

To summarize, there is a need for methods and tools for extracting repeatedly visited personal places and deriving the place semantics (i.e., meanings or purposes) from episodic mobility data of a large number of people. To fulfil this need, several problems must be solved:

- Work effectively with massive, long-term data from a large number of people.
- Be able to deal with the episodic character of the data.
- Respect personal privacy by using only aggregated data for the whole population or large groups of people and avoiding access to individual data, in particular, people's locations.

Finding efficient and privacy-preserving approaches to deriving information equivalent to daily mobility diaries from semantically poor episodic movement data was one of the objectives of an EU-funded research project DataSim (http://www.uhasselt.be/UH/About-DATASIM/Problems-addressed.html). Our research on determining place semantics with the help of visual analytics techniques was a part of this project. The goal was not to develop ready-to-use tools for city planners and transportation managers but to devise a suitable methodology and test its feasibility.

This paper describes the methodology we have devised and demonstrates its feasibility using two case studies with simulated tracks from the VAST Challenge 2014, for which ground truth is available, and real traces built from georeferenced tweets posted during one year within a metropolitan area encompassing San Diego (USA) and the surrounding communities.

The Web site geoanalytics.net/and/papers/placeSemantics contains supplementary materials, which include texts providing additional details and explanations concerning the analytical procedure, a video demonstrating the process of inferring place semantics and the interactive visual techniques employed, and a link to the data we used in this paper.

1 RELATED WORKS

1.1 Obtaining significant places from movement data

The term 'significant places' is henceforth used to refer to places repeatedly and purposefully visited by individuals and therefore having certain meanings (semantics) for the individuals, such as home, work, place for shopping, recreation, social activities, etc.

Potentially interesting places have been extracted from mobility data using different variants of clustering [7][8][9]. Semantic information can be attached to these places by matching them with locations of predefined places of interest (POI) [10][11]. This approach, however, does not uncover personal POIs such as home, work, child's school, and place of regular shopping. Mobile phone data have been used for inferring land use categories based on an observation that residential, commercial, industrial, and green areas significantly differ in temporal profiles of mobile phone activities [12]. A few works describe ad hoc approaches to finding and interpreting personal places. Ahas et al. [13] use mobile phone data to identify home and work places based on the frequency of the person's calls from each place, their average time of the day, and the standard deviation of the time of the day using a set of classification rules that are based on the researchers' background knowledge. Isaacman et al. [14] derived place classification rules by

analysing labelled data of volunteers. The distinguishing attributes were the visit frequency, duration, and the counts of phone use events during work hours (13:00-17:00) and during home hours (19:00-7:00). So far, no systematic approaches to identification of different types of personal POI have been reported.

Mobile phone data have also been used for analysing personal daily mobility behaviours without trying to establish place meanings. Daily sequences of persons' positions can be transformed into graphs with nodes corresponding to visited places and directed links representing trips between the places. Surprisingly, out of millions of possible graph structures, only 17 are of statistical significance. These 17 graphs are called human mobility motifs [15]. Based on this work, it was shown that daily travel behaviours are strikingly similar for different cities and countries while there are also differences explainable by demographic differences between cities [16].

Twitter data, specifically, Foursquare check-in posts extracted from Twitter, were also analysed for deriving location types. It was found that different types of public places (coffee shops, sub shops, book stores, etc.) have characteristic daily and weekly temporal patterns of visits, which can be used for place type classification [17][18]. Better results can be achieved through additional analysis of term occurrences in the text content of the check-in tweets [19].

1.2 Related works in visual analytics

In visual analytics, no specific methods for extraction of semantic information from mobility data have been proposed so far. There are works in which semantic information is derived from contents of georeferenced Twitter or Twitter-like messages and used in spatio-temporal analysis. In VAST Challenge 2011 [20], synthetic Twitter-like data were analysed for understanding the cause, character, and spatio-temporal evolution of a simulated epidemic outbreak; however, these works were not related to place semantics.

Scatterblogs 2 [21] is a real-time monitoring visual analysis tool for Twitter streams using visual composition of content filters to detect and visualize the spatio-temporal distributions of thematic events but not their association with specific places.

Krüger et al. [22] propose to semantically enrich vehicle GPS trajectories with data about public POI obtained from Foursquare by associating POI categories with spatial clusters of trajectory endpoints, i.e., frequent destinations. Their work focuses on visual exploration of trip purposes while accounting for POI categorization uncertainties at multiple spatial scales in space and time.

1.3 Protection of personal privacy

Application of visual analytics methods to data including people's geographic positions is associated with privacy issues [23][24][25]. Aggregation of mobility data of multiple individuals based on spatial and spatio-temporal generalization helps protect personal privacy [26], but this approach is suited for analysis of mass movements rather than personal mobility. Another approach is obfuscation of individuals' real positions, i.e., replacing points by regions; however, obfuscation-based techniques fail to hide repeatedly visited locations and regularly followed routes [23]. The inherent problems of location privacy disappear when mobility data are transformed from geographic space to abstract spaces [5] (section 9.6). There is an example of applying such a transformation to data with previously attached semantic labels [27], but it is not discussed how the required semantic information can be obtained without compromising personal privacy.

Dasgupta and Kosara [28] discuss privacy issues associated with non-geographic data and propose privacy-preserving visualization with parallel coordinates incorporating the formal concepts of *k*-anonymity and *l*-diversity developed in the field of data mining. Our work does not aim at developing a mechanism that formally guarantees chosen levels of privacy protection. The goal is to investigate the principal possibility of extracting personal places and attaching semantic labels to them without accessing geographical positions of individuals.

2 Input Data, Problem Statement, and Methodology Overview

2.1 Format and general characteristics of input data

Our methodology is intended for analysis of episodic human mobility traces, such as records about the use of mobile phones or other mobile devices at various locations. The main components of such data are person's (user's) identifier, location specification, and time stamp. The location specification may be available in the form of geographic coordinates (longitude and latitude) or as a reference to some spatial object with known coordinates, such as mobile network antennas. An example of episodic mobility data we deal with is given in Fig.1, which demonstrates the typical temporal sparseness and irregularity of such data.

UserID	Longitude	Latitude	time
14333974	-117.18329	32.87769	29/09/2012 11:43:18
14333974	-116.8711	32.882217	29/09/2012 13:23:26
14333974	-116.8711	32.882217	29/09/2012 13:49:51
14333974	-116.8711	32.882217	29/09/2012 14:03:43
14333974	-116.8711	32.882217	29/09/2012 14:16:43
14333974	-117.10787	32.921898	29/09/2012 18:36:01
14333974	-117.19572	32.89569	03/10/2012 13:02:37
14333974	-117.196075	32.895653	03/10/2012 13:55:41
14333974	-117.20196	32.896408	03/10/2012 16:56:31
14333974	-117.20266	32.893803	05/10/2012 16:21:21
14333974	-117.09867	32.935722	07/10/2012 10:46:19
14333974	-117.09692	32.93642	07/10/2012 18:29:22

Fig. 1. An example of episodic mobility data.

Mobility data may also have other components, for example, the type of activity, such as phone calling, sending SMS, accessing a web site, etc. Geolocated social media posts include some content (text, photo, or video), which may be supplied with a title and/or hashtags, and media-specific attributes. Although some of components may contain useful location-related information, such as venue types in Foursquare check-in records [17][18][19][22], we intentionally limit the core of our methodology to using only the main components (i.e., person identifier, location, and time), which are present in all kinds of mobility data. This means that the approach does not require any additional components to be present in the data, but it also does not prohibit the use of additional components when available.

The extraction of significant personal places and derivation of the place meanings requires the data to cover a sufficiently long time period and include a sufficient number of presence records for these places. We estimate the shortest meaningful time period to be at least one week for identifying the home and work places and at least two weeks for identifying other types of

places, which are not expected to be visited every day. However, this only applies to data with high temporal frequency, which can come from active mobile phone users or active bloggers. When data are temporally sparse, it is necessary to have records from a much longer time period.

Besides mobility data, the methodology requires additional data that can be used for validating the plausibility of the assignment of semantic labels to places. One possibility is land use (LU) data. For example, when a set of places is going to be labelled as 'home', it should be checked whether most of them are in residential areas. Another possibility is to utilise data about relevant geographical objects, such as public transportation stops, schools, shops, and restaurants. Such objects are commonly called "points of interest" (POI). POI data can be obtained from map feature services, such as OpenStreetMap (www.openstreetmap.org), or retrieved from geographic databases. It is then possible to derive counts of different types of POIs inside places or within a specified distance threshold. Before assigning some meaning to a set of places, the compatibility of this meaning with the POI types occurring in these places should be checked.

Hence, the proposed methodology for acquisition of place semantics requires two datasets: (1) time-stamped positions of people and (2) LU or POI data that can be used for validation of the place meaning assignments.

2.2 Data examples

One of the types of mobility data addressed by our approach is mobile phone use records; however, we have no possibility of using a dataset of this type for published work due to privacy restrictions. Instead, we demonstrate our approach using two openly accessible examples of mobility data, which have properties similar to those of mobile phone use data.

2.2.1 VAST Challenge 2014 Mini-Challenge 2

The original dataset provided for the VAST Challenge 2014 Mini-Challenge 2 [29] consists of simulated tracks of cars with duration of two weeks. The records include timestamps, car identifiers, and coordinates. We used the tracks of 35 personal cars and ignored the tracks of 5 remaining vehicles utilized only for business purposes. The full tracks cannot be classified as episodic mobility data because of high temporal resolution (1 second), which allows determining the exact times of arriving at each visited place and leaving it. To have a suitable example of episodic data, we extracted a subset of the car position records by selecting the positions of stops and taking a 25% sample of these positions. This gave us 1,469 records imitating the properties of episodic mobility data, as depicted in Fig.1.

No data similar to land use or POI data were provided for the challenge. As the underlying territory for the car tacks is fictitious, existing databases or map feature services cannot give us suitable information about places. To create a substitute for POI data, we utilized simulated credit card transaction records, also available for the challenge. The details of pre-processing the VAST Challenge data are given in Appendix I at geoanalytics.net/and/papers/placeSemantics.

We would like to stress that, although the conditions of the challenge did not require it, we analysed the data in a privacy-preserving way, i.e., without looking at any personal data.

2.2.2 San Diego tweets

Georeferenced Twitter messages posted in the metropolitan area of San Diego (USA) were collected over a period of 302 days from the end of September 2012 until the end of July 2013.

In the context of our task, we are only interested in data from residents of the area. To separate residents from visitors, we used a simple filter: there must be at least 100 tweets from a person, and the time span between the first and the last tweets must be at least 100 days. This rather arbitrary filter was nevertheless adequate for obtaining a test dataset; it was not our goal to precisely determine all residents of the area. The selected subset consists of about 2.5 million records of 4,286 individuals. The geographical extent of the area is 112 ×103 km.

To validate place meaning assignments, we obtained land use data for the San Diego area in the form of polygons with labels specifying the land use classes within the polygon boundaries.

Please note that the texts of the tweets were not used as sources of place-related semantic information in this case study. We intentionally used only the minimum subset of fields shared by all kinds of mobility data, i.e., person's identifier, geographic coordinates, and time, to insure the general applicability of the methodology. Still, our approach as such does not preclude the use of information derived from Twitter texts. In our earlier work [30], we analysed the temporal and spatial distributions of occurrences of different topics (subjects) people tweeted about, such as "family", "home", "work", "education", "friends", "food", etc. High frequencies of certain topic(s) in a place may be related to the place meaning; however, it should be taken into account that people may tweet about any topic from any kind of place. For example, one may tweet about work-related topics while being at home or at a beach and tweet about food and drinks while being at work. Hence, the topics occurring in a place cannot be used as absolute indicators of the place meanings but rather as supporting evidence. Our methodology allows the analyst to utilise topic frequencies extracted from message texts in a way similar to the utilisation of land use or POI information. In particular, the multi-attribute bar chart display described in section 4 can be used to visualise and analyse aggregated topic frequencies for sets of places.

2.3 Problem statement

Having episodic mobility data as shown in Fig.1, we aim at obtaining so called "semantic trajectories", which may look as shown in Fig.2. In semantic trajectories, the geographic locations are substituted by semantic labels denoting the meanings of the visited places or types of activities performed. The transformation of the original mobility data into semantic trajectories needs to be done for a large set of individuals in such a way that the geographic positions of the individuals are hidden from the analyst. Please note that the resulting semantic trajectories are devoid of geographic positions and thus can be viewed and further analysed without compromising individual's location privacy.

UserID	Trip N	Trip date	Type of place/activity	Time
14333974	1	29/09/2012	shopping	29/09/2012 11:43:18
14333974	1	29/09/2012	eating	29/09/2012 13:23:26
14333974	1	29/09/2012	eating	29/09/2012 13:49:51
14333974	1	29/09/2012	eating	29/09/2012 14:03:43
14333974	1	29/09/2012	eating	29/09/2012 14:16:43
14333974	1	29/09/2012	outdoor recreation	29/09/2012 18:36:01
14333974	2	03/10/2012	work or study	03/10/2012 13:02:37
14333974	2	03/10/2012	work or study	03/10/2012 13:55:41
14333974	2	03/10/2012	fitness	03/10/2012 16:56:31
14333974	3	07/10/2012	shopping	07/10/2012 10:46:19
14333974	3	07/10/2012	home	07/10/2012 18:29:22

Fig. 2. An example of "semantic trajectories" obtained from episodic mobility data.

To obtain semantic trajectories from mobility data, the following tasks need to be fulfilled:

1. Extract personal places repeatedly visited by each individual.
2. Identify the most likely individual-specific meanings of the extracted personal places and attach semantic labels denoting these meanings to the places.
3. For the entire set of individuals, extract public places visited by multiple individuals.
4. Identify the most likely meanings of the public places and attach corresponding semantic labels to the places.
5. For each point in the original data, find a personal or public place containing it and replace the geographic reference of the point by the semantic label of the place.

The resulting semantic trajectories do not yet adequately approximate personal mobility diaries because they are episodic, like the original mobility data. That is, the trajectories include only a subset of places that were actually visited on each day. For example, trip 1 in Fig. 2 begins with 'shopping' at 11:43, although the person was, most probably, at home in the morning and could have also visited other places before appearing in the shopping place. It can also be guessed that the person returned home after 'outdoor recreation', but this is not reflected in the data. Hence, in the next stage of analysis, more complete daily semantic trajectories should be reconstructed from partial semantic trajectories. This next stage is, however, beyond the scope of this paper.

2.4 Methodology overview

To support tasks 1-5 and following analysis of semantic trajectories, we have devised an analytical workflow shown schematically in Fig. 3. On the left, the workflow is represented by a flow chart. In the centre, there are brief comments to the steps of the procedure. On the right, there are references to the paper sections and Appendix II where these steps are described.

For extracting personal and public places, we developed an automated tool involving clustering of points from episodic mobility data by spatial proximity. To extract personal places, the points of each individual are clustered separately; to extract public places, the points of all people are clustered together. Places are defined by constructing boundaries (spatial convex hulls or buffers) around the point clusters. The tool works automatically. It takes input data from the database, processes them, and puts the resulting place boundaries back in the database. Personal places are associated with the identifiers of the individuals they belong to; however, the places and the identifiers of their owners are not shown to the analyst. The place extraction algorithm is described in detail in section 3.

A suite of interactive visual techniques supports the process of determining place meanings. To avoid disclosing possibly sensitive personal information, we have chosen such visualization methods that only show data aggregated over the whole set or groups (clusters) of places and allow no access to individual data:

- *multi-attribute summary bar chart* showing value summaries of multiple numeric attributes (section 4);
- *qualitative histogram* showing aggregated counts of qualitative values (section 4);
- *two-dimensional (2d) time histogram* showing aggregated two-dimensional time series of place visits (section 5).

Identifying the most probable home and work places is supported by a place ranking tool described in section 6.1. The places of each person are ranked based on several relevant attributes, such as the total number of visit-days and the proportions of visits in the typical work and home hours. The places with the best ranks are considered as candidates for receiving the target meaning, i.e., ‘home’ or ‘work’. To check if these places are sufficiently good candidates, the analyst looks at the associated LU or POI classes, as described in section 6.2. The analyst investigates how modifications of the criteria weights affect the selection of the candidate places and the corresponding statistics of the LU or POI classes. Finally, the analyst selects the best set of candidates and assigns the target meaning to them. For target meanings other than ‘home’ and ‘work’, candidate places are selected by means of interactive filtering based on relevant temporal attributes and land use or POI classes. A detailed illustrated example of inferring place meanings from the VAST Challenge data is given in Appendix III and in an accompanying video, and Section 7 contains general guidelines for place semantics acquisition.

After assigning semantic labels to personal and public places, the derivation of semantic trajectories is performed in a straightforward way (see section 2.3, task 5). The resulting semantic trajectories can be explored using a map of an abstract semantic space, as described in section 8. In particular, the plausibility of the place meaning assignments can be checked by analysing the emerging patterns of flows (aggregate movements) in the semantic space.

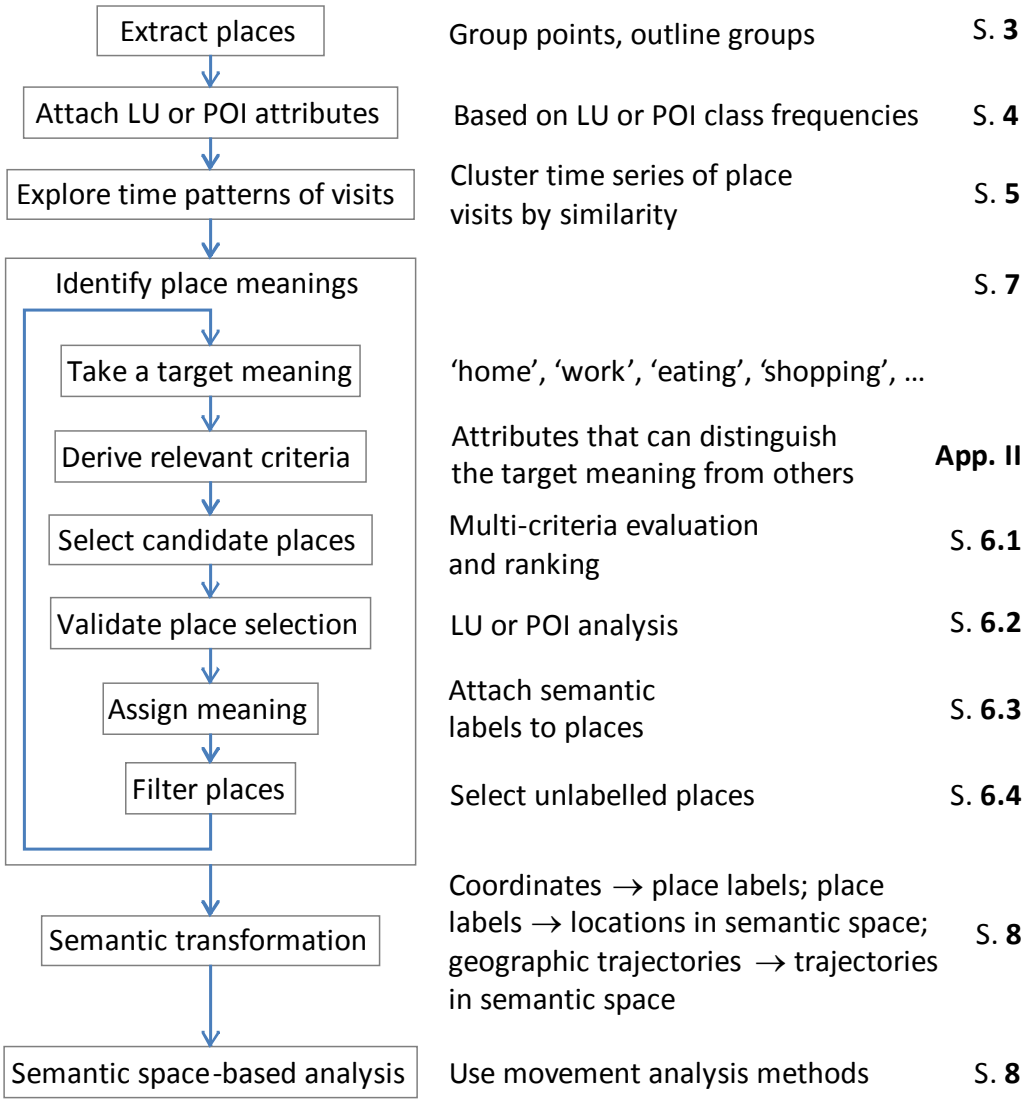

Fig. 3. The analytical workflow for extracting semantic information from mobility data and subsequent semantic analysis.

3 EXTRACTION OF PERSONAL AND PUBLIC PLACES

In mobility data, the positions of moving objects are often specified as points in the geographic space. Starting from these points, it is necessary to find and delineate repeatedly visited places of each person. Places of interest can be obtained from point data by clustering points in space and building spatial buffers or convex hulls around the clusters [31]. In our case, clustering of points and place delineation needed to be done separately for each person. Density-based clustering used in previous research [31] is not fully suitable for the task of extracting personal places. A property of density-based clustering algorithms is that they can build clusters of arbitrary shapes and sizes; thus, they can easily construct a huge cluster covering the whole city centre if the point density is sufficiently high throughout the area. Such a cluster would represent not a single place but multiple places, which is not desirable. Smaller density-based clusters can be obtained by increasing the density threshold. The problem is that the same threshold is applied throughout the

whole study space. Since point concentrations are usually not equally dense in different parts of the study area, clustering with high density threshold will miss many point concentrations with lower densities while clustering with low density threshold will merge together multiple point concentrations located close to one another. Moreover, the density of point concentrations varies not only across space but also across the population: it depends on how actively a person tweets or uses a mobile phone. It is both infeasible and violating personal privacy to look at the point distribution of each person in order to select an individual-based density threshold. The clustering and place delineation must be done in batch mode for all persons; hence, the same parameters have to be used.

We have established thus far that the task of personal place extraction requires a point clustering algorithm that is insensitive to the density variation and allows limiting the spatial extents of the resulting clusters. A spatially bounded point clustering algorithm, used earlier for generating space tessellations [32], can be adapted for this purpose. In short, the algorithm places points in circles with a user-specified maximal radius R_{max}. When a point is added to a circle, the circle centre is re-computed by averaging the x- and y-coordinates of all its points. When there is no suitable circle for a point, a new circle with the centre at this point is created. After processing all points, the circles containing fewer points than the user-chosen minimal number are discarded, and spatial clusters are formed from the points of the remaining circles. The algorithm allows different point densities in different circles and does not allow the clusters to grow beyond the specified limit R_{max}. Please note that the resulting clusters only consist of the points and do not include the enclosing circles; hence, the clusters may have smaller radii than R_{max} and may have arbitrary shapes.

The same algorithm may be used for extracting public places, i.e., places visited by many individuals. The difference from extracting personal places is that the algorithm is applied to points representing locations of all persons at once. After the algorithm finishes, only clusters containing points of at least a given minimum number of different persons are retained as the result.

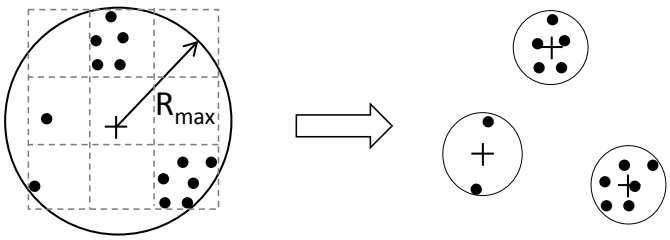

Fig. 4. A group of points where the maximal density is not attained near the geometric centre is subdivided into smaller point groups.

Fig. 5. Neighbouring point clusters may be united into larger clusters.

A drawback of this algorithm is that some point groupings may not look like "true" spatial clusters. Thus, two or three spatially compact concentrations of points may be grouped together if they fit in one circle, or a concentration of points may be united with one or a few isolated points scattered around. An illustration of this case is given in Fig. 4, left. In such artificial groupings, the maximal point density is not attained at the geometric cluster centre but close to the periphery. This can be used as an indication of poor grouping. Unnatural groupings may be not a problem for generating spatial tessellations, which has been the original goal of the algorithm. However, for the purpose of extracting significant personal or public places, such groupings should be avoided. To alleviate this problem, we have modified the algorithm in the following way.

Input: set of points $P = \{(x,y)\}$.

Parameters: maximal and minimal radii R_{max} and R_{min}; minimal number of points in a group N_{min}.

Algorithm:

1. Apply the base point clustering algorithm to P with the parameter value R_{max}. Let Γ be the list of resulting groups of points.
2. Go through Γ. For each group of points $G \in \Gamma$ do the following:
   2.1 Let R be the group radius. If $R<R_{min}$, go to the next group.
   2.2 Else, build a bounding box containing G and divide it into 9 equal rectangles by two horizontal and two vertical lines (Fig. 1 left). Check if the central rectangle contains fewer points than any other rectangle (this means that the maximal point density is not attained near the geometric centre of the cluster). If not, go to the next group.
   2.3 Else, subdivide G into smaller groups (Fig. 4 right) by applying the base algorithm to G with the parameter value R/2. Add the resulting groups to Γ and remove G from Γ.
   2.4 Redistribute points from the neighbouring groups: if a point is closer to the centre of one of the new groups than to the centre of its current group, move it to the new group.
3. Remove from Γ the groups where the number of points is less than N_{min}. Return Γ as the result.

For extracting personal places from the San Diego data, we used the following parameters: R_{max}=150m, R_{min}=75m, N_{min}=5. The rationale for choosing the radius range 75-150 m is as follows. We want to avoid combining points related to several semantically different places into one cluster; hence, the radius limit should be small. But then, we should account for possible

position errors in the data. Particularly, mobile devices often determine positions using WiFi, which is less accurate than GPS. Empirical studies [33][34] have shown that WiFi positioning errors may be much larger than the 20 – 40 m officially reported. In the experiments, 70-90% of the errors were within the range of 0-150 m, and the median error was about 74m. Hence, using a value smaller than 150 m as the maximal cluster radius is not advisable.

The VAST Challenge data are supposed to imitate GPS data, which are usually quite accurate; however, the data providers intentionally introduced noise in some tracks. We have taken R_{max}=100m, R_{min}=75m, N_{min}=2. The latter value is small because the time span of the data is only two weeks; even a place that was visited only twice in two weeks may be one of regularly visited personal places.

Another problem to deal with in place extraction is that some places may be quite large. For example, a person may work or study on the campus of University of California, San Diego, the extent of which is 3×2 km. The person's points located on the campus may form multiple small clusters (Fig. 5A; points are represented by pink semi-transparent circles), which will be undesirably interpreted as separate personal places. When the person's visits are aggregated into such small places, the resulting temporal patterns for these places might not resemble a typical temporal pattern of attending a place of work or study. Please note that points of one person are shown in Fig. 5 for illustration purpose only. Analysing data of individuals is not a part of the suggested methodology.

Hence, it is desirable to unite neighbouring small clusters into larger clusters while avoiding the weaknesses of density-based algorithms discussed earlier. To do this, we further extend the space-bounded point clustering algorithm on the basis of the following observation: when two clusters are very close in space, there are points in each of them that could also be the members of the other cluster, i.e., their distances to the centre of the other cluster are below the maximal circle radius. We call such points "connecting points". The idea is to unite clusters sharing at least a chosen minimal number of connecting points C_{min}. To avoid extreme growth of the resulting clusters, the user may set an upper limit R_{max}^{+} on the radius of merged clusters. Two clusters are not merged if this would result in exceeding R_{max}^{+}. The following algorithm extension is proposed:

1. For each pair of neighbouring clusters, find the number of connecting points. Neighbouring clusters are found using the spatial index, which is a part of the base algorithm [32]. Make a list of cluster pairs having at least C_{min} connecting points.
2. Sort the list of connected cluster pairs in the order of decreasing number of connecting points.
3. Go through the sorted list. For each pair do the following steps:
    3.1 If the radius of the circle enclosing the two clusters exceeds R_{max}^{+}, skip this pair and go to the next one.
    3.2 Else, merge the two clusters. Replace the occurrences of the identifiers of the original clusters in the sorted list by the identifier of the merged cluster. Go to the next pair.

According to this algorithm, strongly connected clusters get higher probability of being merged than more loosely connected clusters.

The impact of the parameter R_{max}^{+} is illustrated in Fig. 5. In Fig. 5B, the blue polygons are the convex hulls of the clusters obtained through merging original clusters connected by at least 3

points without limiting the extents of the resulting clusters. In Fig. 5C, the green polygons are the convex hulls of the clusters obtained with R_{max}^{+} = 600 m.

For assessing the appropriateness of the clustering results, the analyst looks at the statistics of the resulting cluster radii. The presence of too large values may require the analyst to look at the outlines of the largest clusters, which are drawn on a map without showing the identifiers of the place owners. The results are acceptable when the largest clusters correspond to large geographic objects, such as university campuses, visible on the background map. The spatial convex hulls of the clusters are taken as the place boundaries.

For the San Diego data, clustering without setting R_{max}^{+} results in obtaining personal places with radii up to 1.5 km and public places with radii up to 4 km, which we judged as too large. We have empirically found that the upper limit of 600 m works well enough. We extracted 38,225 personal and 9,301 public places. For the artificial data from the VAST Challenge, there is no possibility to check the largest places against a real geographical background. We set R_{max}^{+} to 250 m, to avoid obtaining too big places. We obtained 202 personal and 41 public places.

4 ATTACHING LAND USE OR POI INFORMATION TO PLACES

To be able to validate place meaning assignments and resolve ambiguities, land use or POI information needs to be attached to the extracted places. For this purpose, our tools derive a frequency distribution of distinct LU or POI classes for each place. This is done differently for LU and POI data. For LU data, which are usually available in the form of polygons labelled by land use classes, the frequency distributions are obtained in the following way. First, for each point in the original data, the containing land use polygon is found, and the class label of this polygon is attached to the point. Second, for each extracted place, the frequencies of occurrence of distinct LU classes among the points contained in this place are counted. For POI data, which consist of place coordinates and labels signifying the POI classes, the procedure is as follows. For each extracted place, all POIs contained within the place boundaries are found, and the frequencies of distinct POI classes are counted.

Depending on the total number N of distinct LU or POI classes occurring on the studied territory, the class frequency information may be stored in two different ways. If N is not large, the information can be represented by N numeric attributes, one for each class. The values of the attributes for each place are the frequencies of the N classes. We used this approach in the VAST Challenge case study, where we had 8 distinct POI classes. In case of large N, the class frequency information can be represented by k qualitative (nominal) attributes, where k is a number chosen by the analyst, such that $1 \le k < N$. The attributes can be named "most frequent class", "second frequent class", …, "k-th frequent class". Their values are the first, second, …, k-th class labels in the label arrangement by the decreasing frequencies. We used this approach in the San Diego case study, where N=103. It would be difficult to deal with 103 numeric attributes representing all possible land use classes. We found it sufficient to use instead three qualitative attributes, i.e., we took k=3.

For viewing LU or POI information and using it in the analysis, we have developed two types of aggregated data displays. For dealing with a set of numeric attributes representing frequencies (in particular, frequencies of different LU or POI classes), we use a multi-attribute summary bar chart display. The examples in Figs. 6 and 7 show aggregated frequencies of different POI classes by proportional lengths of horizontal bars. There are interactive controls for selecting the

aggregation operation (sum, average, minimum, maximum, or count) and setting the value filtering condition. The display in Fig. 6 applies the operation 'average' and shows the average percentages of different POI classes per place. The value filtering condition is '>0'. The display in Fig. 7 applies the operation 'count' and the value filtering condition '>=5', i.e., it shows, for each POI class, the counts of places where at least 5% of the points have this POI class.

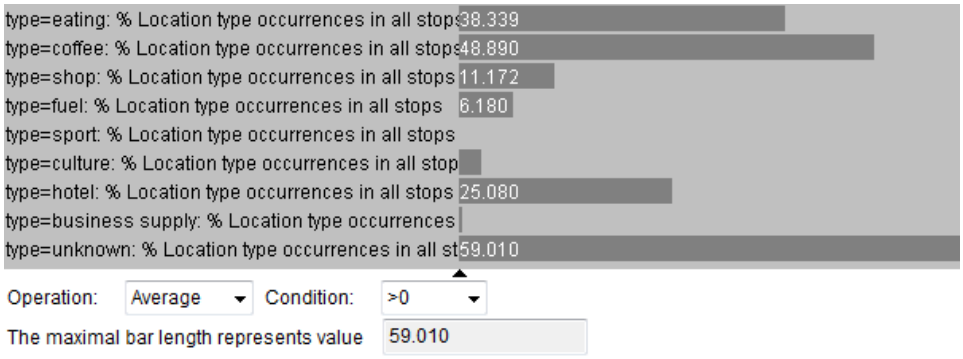

Fig. 6. A multi-attribute summary bar chart of POI class frequencies in the places extracted from the VAST Challenge data.

To compare LU or POI class frequencies in two or more groups of places, a colour propagation mechanism is used. When places are divided into groups in any way, such as clustering or classification according to the likelihood of having a certain meaning, distinct colours are assigned to the groups. Information about the group colours and the group membership of each place can be propagated to all currently existing displays. A multi-attribute bar chart reacts to the colour propagation by multiplying the bars: it creates as many groups of bars as there are groups of places. Each group of bars represents aggregated information for one group of places; the bars are painted in the colour of this group. This is illustrated in Fig. 7. Where the deviations of the group aggregates from the whole set aggregates are statistically significant, the bars are enclosed in black or white frames indicating significantly higher or significantly lower values, respectively. The significance of the deviations is determined using the chi-square test.

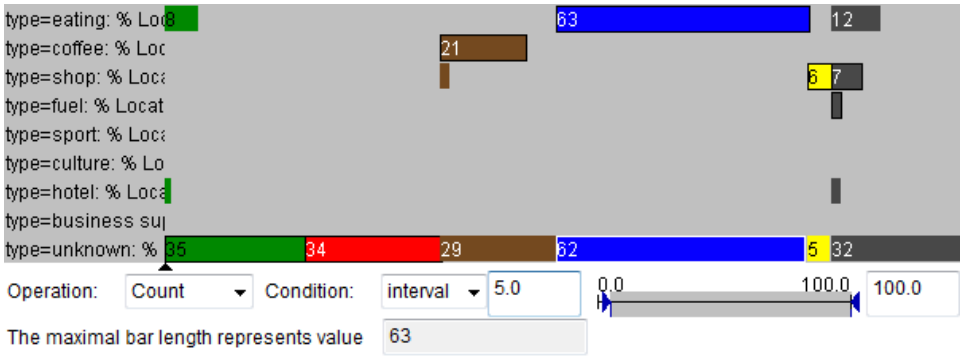

Fig. 7. A multi-attribute summary bar chart shows POI class frequencies for groups of places.

For dealing with qualitative attributes, such as the first, second, …, *k*-th most frequent LU/POI class, we designed a qualitative histogram display. When an attribute has a large number of distinct values, the values can be organised in a hierarchy. For the San Diego case study, where there are 103 different land use classes, we created a hierarchy that can be utilised by the display.

Fig. 8 demonstrates a hierarchical qualitative histogram for the attribute "Land use: most frequent value" of the personal places extracted from the San Diego data. The bar lengths are proportional to the counts of the classes and generalised categories of land use. The upmost bar corresponds to all land use classes taken together. The display reacts to the propagation of place groups and their colours in the same way as shown in Fig. 7.

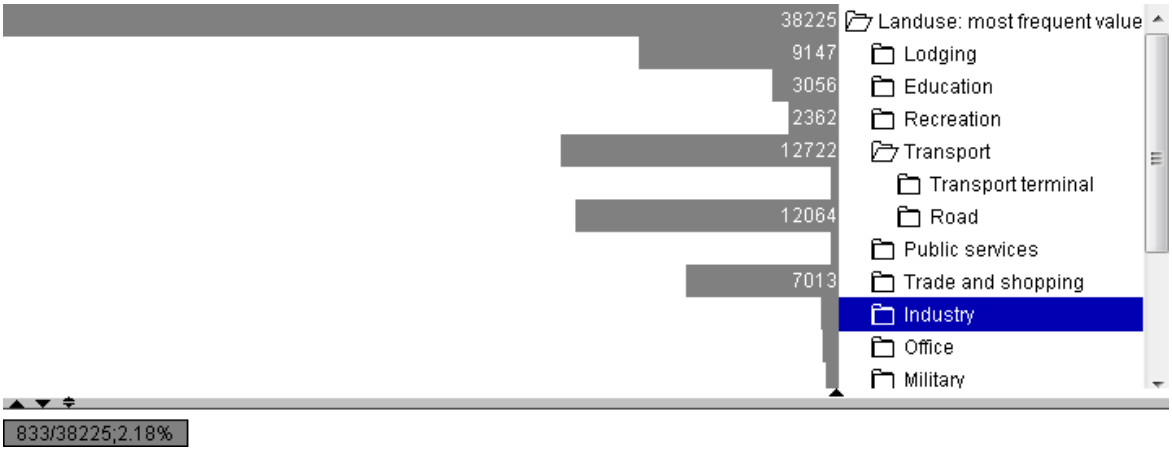

Fig. 8. A hierarchical qualitative histogram shows counts of places with different land use classes for the whole set of personal places extracted from the San Diego data.

5 EXPLORATION OF TIME PATTERNS OF PLACE VISITS

The place extraction tool described in Section 3 computes for each place the total number of visit-days and a two-dimensional time series of place visits by days of the week and hours of the day. For personal places, only the visits of the place owners are counted. Counts of visits are not the same as counts of points. If two consecutive points of a person fit in the same place and the same hour, they are treated as representing the same visit.

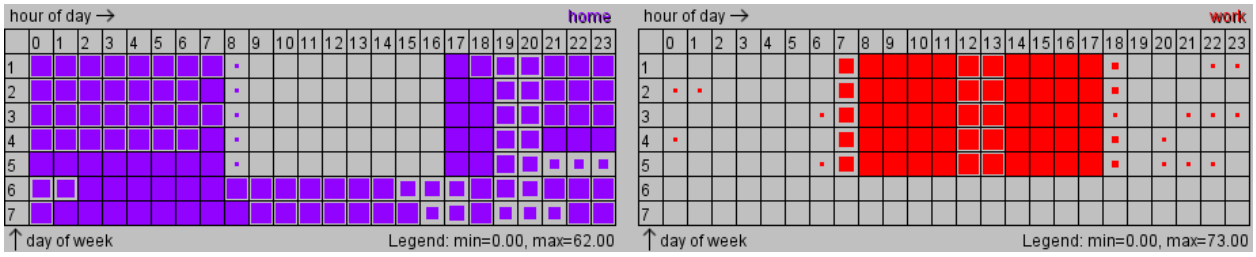

Fig. 9. Based on these temporal patterns of place visits, the places can be interpreted as home (left) and work (right).

For looking at the place visit time series and using them in the analysis, we have designed a two-dimensional time histogram display illustrated in Fig. 9. A time histogram is a matrix with rows corresponding to days of the week and columns to hours of the day. Inside the cells, there are symbols with sizes proportional to aggregated visits in the corresponding days and hours for the set of places currently considered. The UI allows the user to select the aggregation operation (sum, minimum, maximum, average, or count), the condition for including attribute values in the aggregates (all, positive, negative, zero, or within a user-specified interval), and the way of representing the aggregates in the matrix cells (squares, circles, vertical bars, or horizontal bars).

The display reacts to propagation of place groups and their colours (as explained in section 4) by multiplying the time histogram, so that one histogram represents the whole set of places and the remaining histograms show aggregated counts for different groups of places. The symbols in the cells of the matrices have the colours of the clusters. When all histograms do not fit in the window, the display can be scrolled.

Figure 9 illustrates the main idea of our approach to identifying place meanings based on the time patterns of the place visits. Here, time series of visits to two different places are depicted. Based on the observed temporal patterns and our background knowledge concerning the typical times of various activities of people, we can assign the meaning ‘home’ to one place and the meaning ‘work’ to the other place.

Since we need to attach meanings (semantic labels) to a large number of places, we cannot do this by separately looking at the temporal visit pattern of each place. A more scalable approach is to filter and rank the places based on relevant summary attributes derived from the time series. An example of such a derived attribute is the number or proportion of place visits fitting in the work times, i.e., in the hours from 05 to 18 during the work days. We have implemented interactive tools for derivation of relevant attributes, which are described in Appendix II.

However, before deriving attributes that can appropriately distinguish possible place meanings, it is useful to perform an initial exploration of the set of the existing time patterns of place visits. The initial exploration can be done by means of clustering of the place visit time series by similarity and analysing the clusters with the use of 2d time histograms. To standardize the time series across the places and people, the absolute counts are converted to percentages of the total number of the place visits. The clustering can be done using any existing clustering algorithm. The illustration in Fig. 10 represents selected clusters obtained with the k-means clustering method for the San Diego example. The appearances of some patterns suggest the likely meanings of the places in the respective clusters. Thus, patterns ‘a’ and ‘b’ in Fig. 10 suggest the meaning ‘home’, patterns ‘c’ and ‘d’ evoke the meaning ‘work or study’, ‘e’ could be a shopping pattern, and ‘f’ may be attributed to social activities, i.e., visiting or meeting friends or relatives. Such observations may give an idea about the possible activities and their typical times for the territory and population under study.

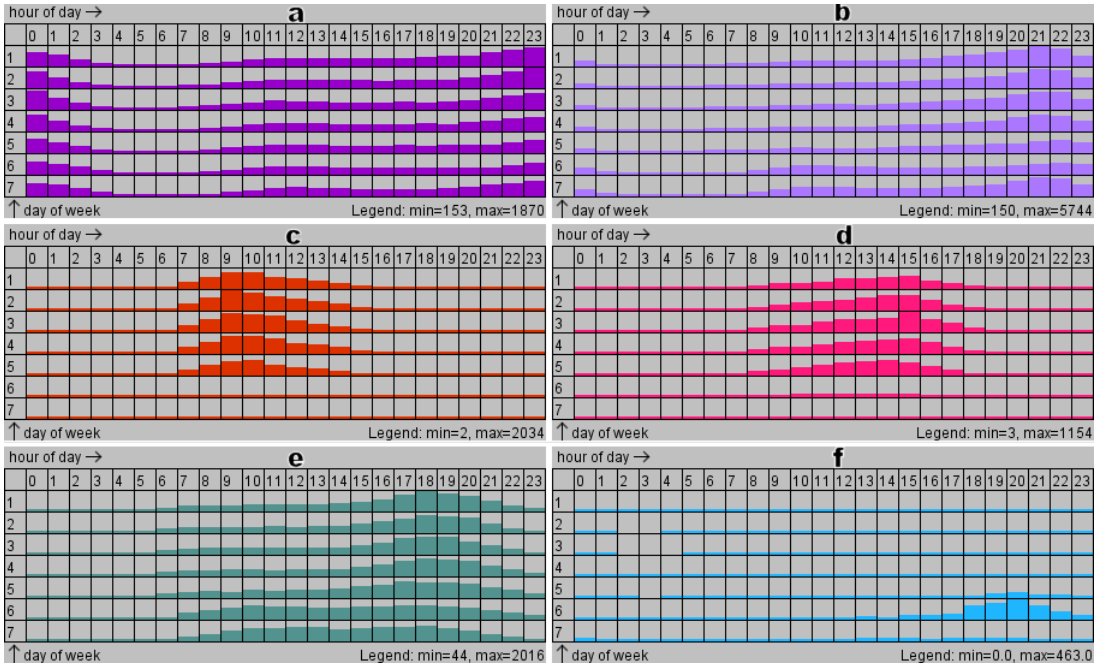

Fig. 10. 2d time histograms for selected clusters of places (San Diego example).

6 INFERRING PLACE MEANINGS

6.1 Multi-criteria evaluation and ranking

From the literature and communication with other researchers, we have discerned several criteria used for identifying home and work places of individuals: place visit frequency, average time of the day when visits happen, counts of visits in the night time and in the work time, and counts of daily trip starts and ends. All these criteria seem relevant, but it has not been known, which one(s) of them works the best. It seems appropriate to combine these criteria, but so far this has only been done in specific studies through ad-hoc rules [13][14].

We have developed a generic interactive technique for combining multiple criteria based on approaches from decision support science [35] where multiple attributes of decision options are integrated into scores representing the degree of suitability or utility of the options. The attributes may be given different weights according to their relative importance. Two types of criteria are distinguished: benefit criteria with higher values being preferred and cost criteria with lower values being preferred. The most common criteria integration method in multi-criteria decision making (MCDM) is the weighted linear combination [36]. The attribute values are normalized to the range [0, 1] depending on their relative positions between the minimal and maximal values. For a benefit criterion, 0 corresponds to the minimum and 1 to the maximum; for a cost criterion, it is the other way around. The integrated score is computed as the sum of the products of the normalized values multiplied by the attribute weights and divided by the sum of the weights. The result is a number ranging from 0 to 1, where 0 is the worst and 1 is the best.

Previously suggested interactive techniques for MCDM [37][38] allow the user to select relevant criteria and set their directionality (cost or benefit) and weights. The evaluation results, i.e., the scores and ranks of the options, are immediately shown on visual displays, such as parallel coordinates plot and map. The user can test how changes of the criteria weights affect the results.

The criteria integration approach from MCDM can be adapted to the task of identifying the most likely home or work place among the personal places of an individual. By nature of the task, the evaluation must be done separately for places of each individual. However, the size of the data and the privacy constraints do not allow separate consideration of the places of each individual by analysts. Analysts can only choose relevant criteria and set a common set of weights to be used for all individuals. A prototypical UI is shown in Fig. 11. A computational tool calculates integrated scores separately for the places of each individual and ranks the places according to their scores. The target meaning (home, work, etc.) can be assigned to the set of the top ranking places of all individuals. We stress that the meaning is assigned to multiple places simultaneously without looking at any particular place, i.e., without accessing personal data.

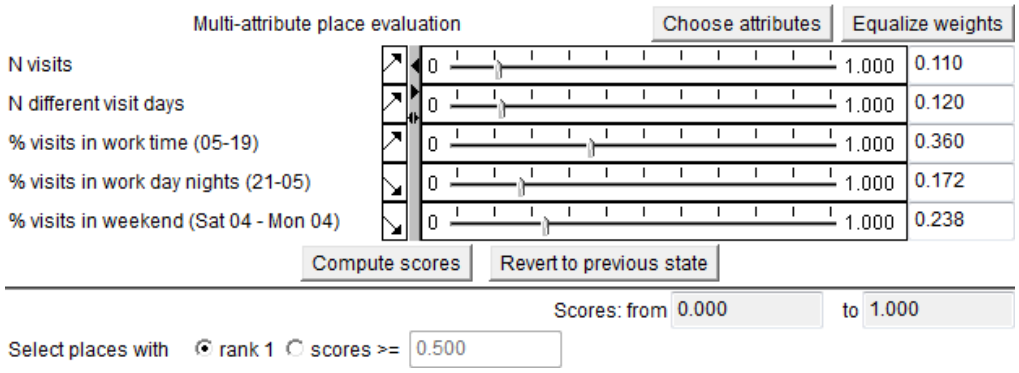

Fig. 11. A fragment of the UI for the multi-criteria interpretation of personal places.

We have developed two multi-criteria ranking tools for personal and public places. The former applies ranking separately to personal places of each individual, whereas the latter applies ranking to all public places at once. The tools have similar UIs and functionalities.

6.2 Validation of place scores and ranks

As a part of multi-criteria evaluation, two problems need to be solved: how to assess the quality of the evaluation outputs and how to test the impact of choosing different criteria and modifying their weights. The usual MCDM support tools involve visualization of characteristics and scores of individual options, e.g., on a parallel coordinates plot [37][38]. This allows the user to check the suitability of one or a few best scoring options by comparing their characteristics with those of the other options. The sensitivity of the evaluation results to the settings can be investigated by interactively changing the settings and observing the consequent changes in the display.

This approach is not applicable to the task of personal place evaluation because it would require separate consideration and comparison of places for each person, which is both infeasible and violates individual privacy. One solution we propose involves land use data. Among the LU classes, there are classes associated with the meaning ‘home’ (various residential land uses), classes related to work or study (industry, education, office, construction, military land use), and classes related to other activities (shopping, recreation, health care, etc.). The quality of place ranking results can be judged from the proportion of the places in land use categories relevant to the target meaning among the best scoring places.

Let us consider the example of ranking the personal places extracted from the San Diego data with regard to the target meaning ‘work or study’. We have chosen the attributes shown in Fig.

11 as the ranking criteria, set their directionality (benefit or cost), and obtained initial ranks with all criteria having equal weights. The ranking tool creates an attribute ‘Best scored’ and attaches values ‘y’ (yes) and ‘n’ (no) to the places. By means of interactive filtering, we select the subset of places that have the value ‘y’ and examine the land use classes occurring in this subset using a qualitative histogram (Fig. 12). We observe that the subset includes a number of places with LU classes relevant to the target meaning ‘work or study’, but there are also many occurrences of irrelevant LU classes. We are going to modify the criteria weights and check whether this improves the proportions of the relevant land use classes.

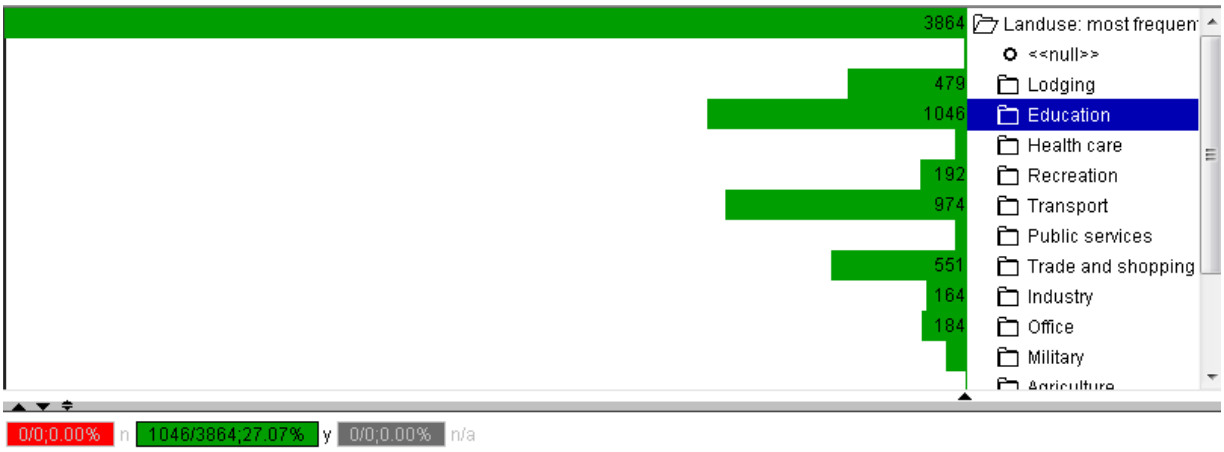

Fig. 12. The distribution of different land use classes across the set of places having the highest ranks with regard to the target meaning ‘work or study’.

To support the comparison between outputs of two consecutive rankings, the evaluation tool creates a special attribute representing the place rank changes. The possible values of the attribute are ‘yy’ (the highest rank in both outputs), ‘nn’ (lower ranks in both outputs), ‘yn’ (the highest rank changed to a lower rank), and ‘ny’ (a lower rank changed to the highest rank). These values define classes of the places, which can be propagated to the land use histogram. The tool remembers the previously used criteria weights, so that the analyst can restore these weights and the corresponding place scores and ranks if the result of the changes is not satisfactory.

Continuing our example, we modify the criteria weights as shown in Fig. 11, re-compute the place ranks, and propagate the place classes expressing the rank changes to the land use histogram. By means of interactive filtering, we focus only on those places that have changed their ranks, i.e., the classes ‘ny’ and ‘yn’. In the histogram (Fig. 13), these classes are represented by the green and orange bars, respectively. We see that the change in the weight values has increased the number of the top-ranked places located in the relevant categories “Education” and “Office”: the respective green bars are longer than the orange ones. The number of the top ranked places located in the irrelevant land use category “Lodging” has decreased. Hence, the change in the weight values has improved the place evaluation result in terms of the relevance of the LU classes. Still, it is also necessary to assess the changes with regard to the temporal patterns of the place visits. We do this using a 2d time histogram display. It shows us (Fig. 14) that the visit patterns of the places that improved their ranks match much better the target meaning ‘work or study’ than the patterns of the places with lowered ranks. Hence, the results of the weight modification can be approved.

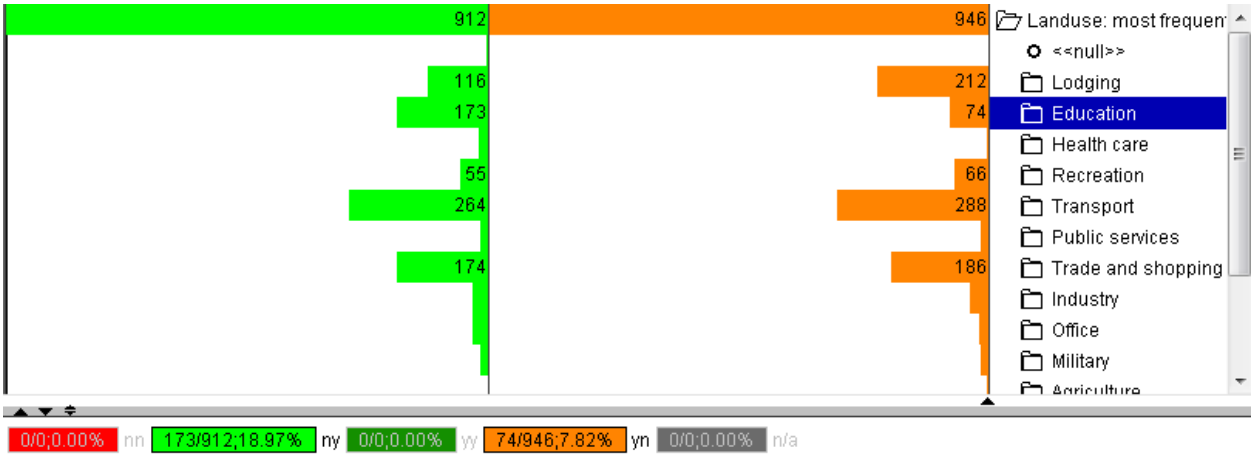

Fig. 13. The qualitative histogram shows aggregate statistics of the land use classes for the places that changed their ranks due to a modification of criteria weights.

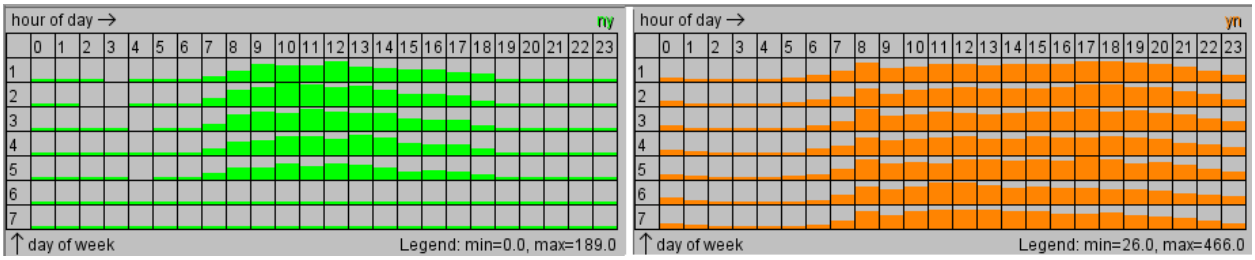

Fig. 14. The 2d time histograms show the aggregate temporal patterns of visits for the places that changed their ranks due to a modification of criteria weights.

Information about the relevance of the LU classes can be transformed into a binary attribute with values 1 and 0 showing whether the list of land use classes attached to a place includes any target-relevant class. This is facilitated by a generic tool creating binary attributes based on object filtering. Binary attributes can be used for a direct investigation of the relationship between the place scores and the land use relevance by means of a *2d cross-histogram* display (Fig. 15). Its two dimensions correspond to two attributes. In Fig. 15, the dimensions are the evaluation score (horizontal dimension) and land use relevance (vertical dimension). The bins correspond to values or value intervals of the attributes. The bars in the bins represent the frequencies of the corresponding value pairs. The cross-histogram can be displayed in a cumulative mode, in which the frequencies are accumulated along the rows or along the columns from the minimal to the maximal value of the respective attribute or in the opposite direction. In Fig. 15, the frequencies are accumulated along the rows in the direction from the maximal score to the minimal; hence, a bar in a score bin $(x_i, x_{i+1}]$ represents the number of places with the scores greater than x_i. In the lower row, it is the number of places with target-irrelevant LU classes, and in the upper row, it is the number of places with target-relevant land uses.

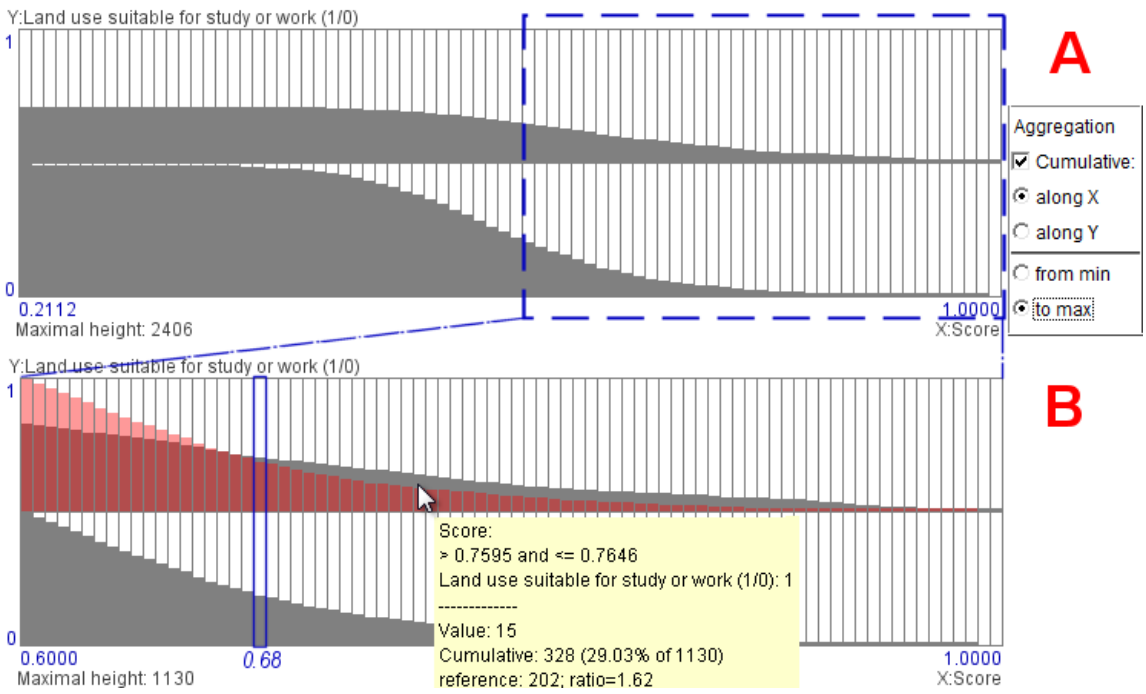

Fig. 15. A 2d cross-histogram shows the relationship between the evaluation scores and the land use relevance to the target meaning.

In Fig. 15A, the cross-histogram represents the whole range of the attribute ‘score’. At the left end of the histogram, the cumulative number of places with relevant land uses (upper row) is much smaller than the number of places with irrelevant land uses (lower row). At about 40% of the score range, the number of places with irrelevant land uses starts to steeply decrease as the score increases while the decline in the upper row is much gentler. At some point, the number of places with relevant land uses starts to exceed the number of places with irrelevant land uses. However, it is hard to compare the bar heights in two rows. In order to support the comparison, the display UI allows superimposing a user-selected reference row or column on all other rows or columns. The superimposed histogram is shown in semi-transparent red. In Fig. 15B, where we have focused on the score interval [0.6, 1], the lower row is superimposed on the upper row. By moving the mouse cursor along the x-axis, we can see that, starting from the score of about 0.68, the number of places with relevant land uses exceeds the number of places with irrelevant land uses, and the ratio between these numbers increases with increasing the score. This kind of dependency indicates that the place evaluation scores represent quite well the place’s likelihood of having the target meaning.

There is a technical possibility of implementing a computational tool that could search for the best combination of criteria weights maximising the absolute number and proportion of the top ranked places with target-relevant land uses (i.e., with value 1 of the relevance attribute). However, the process of finding corresponding places for a given target meaning cannot be fully automated. The analyst would need to check, first, if the automatically selected weights corresponded to the analyst’s understanding of the relative importance of the criteria; second, if the distribution of the land use classes was good enough; and, third, if the temporal patterns of the visits to the top ranked places were plausible for the target meaning. For these tasks, the analyst needs interactive visual tools. After the inspection, the analyst may decide to try adding

new criteria. Such decisions and subsequent choices including a possibility of incorporating additional relevant criteria cannot be done automatically.

Despite being important for the validation of place evaluation and ranking, land use data need to be treated with caution. Due to spatial positioning errors, the land use classes of the original data points obtained by calculating their containment in land use polygons may be erroneous. Besides, land use data can be outdated. Hence, land use information may be utilized as supporting evidence but not as a decisive criterion for determining place meanings.

Another approach to place ranking validation and parameter sensitivity testing involves POI data. The difference with regard to land use data is that there may be no specific POI types related to the target meaning 'home'. In this situation, places can be considered as good candidates for the target meaning 'home' when they contain none or very few public POIs. An example of inferring place meanings with the use of POI data is provided in Appendix III and in the accompanying video available at geoanalytics.net/and/papers/placeSemantics.

6.3 Place meaning assignment

An interactive interface for place meaning assignment includes controls for assigning meanings to places based on (a) their ranks, (b) their scores, or (c) current filter. The analyst provides an arbitrary textual label representing a meaning, for instance, "home", "work or study", "shopping", etc. When option (a) is chosen, the meaning is attached to the top ranking places. With option (b), the meaning is attached to the places with the scores above a user-provided threshold. A suitable threshold can be chosen using a 2d cross-histogram display as shown in Fig. 15. Using option (c), the meaning is attached to the places satisfying currently operating filter, irrespective of the place scores or ranks. Options (b) and (c) are mostly used for public places and for personal places when the target meaning is not 'home' or 'work', i.e., when the number of personal places that may be expected to have this target meaning is not limited. For example, there may be multiple shops or restaurants repeatedly visited by an individual.

6.4 Interactive filtering

The analyst can interactively create filters to select subsets of individuals and subsets of places. Various types of interactive filters are described in book [5] (section 4.2). Several filters of diverse types can be combined. For the task of place meaning discovery, the most important are the attribute-based filter, the class/cluster-based filter, and the related set filter. The *attribute-based filter* selects objects based on the values of one or more attributes. The *class/cluster-based filter* selects classes or clusters of objects. The *related set filter* propagates filtering from object set A to another object set B or in the opposite direction when the objects in set A have references to objects of set B, that is, the values of some attribute of the objects in A are identifiers of objects from B. Filter propagation means that the analyst selects a subset of one of the sets by any filter, and the related set filter selects only those objects from the other set that are related to the selected objects of the former set.

In the context of our task, there is a set of personal places and a set of place owners; the places have references to their owners. The analyst can select, for instance, the individuals who have yet been assigned a place with meaning 'home' and use the related set filter to select only the places of these individuals for further analysis. Conversely, the analyst can select the places with the meaning 'work or study' and then select the owners of these places to see how many of them

already have places with the meaning ‘home’. Please note that filtering results are always represented in an aggregated form, to preclude any access to possibly sensitive individual data.

7 GENERAL GUIDELINES FOR INFERRING PLACE SEMANTICS

As it can be seen from the examples discussed above, determining place meanings is an explorative activity strongly relying on analyst’s reasoning, surmises, and insights and often involving trials and errors. It is hardly possible to describe it in the form of an algorithm. Still, based on our experience, we can propose several general guidelines, which refer to the box ‘Identify place meanings’ of the flow chart in Fig. 3.

1. For each target meaning (‘home’, ‘work’, ‘shopping’, etc.), determine relevant attributes, by which places with this meaning can be distinguished from places with other meanings. If needed, derive attributes that are not initially available (see Appendix II).
2. Apply filtering based on the relevant attributes to select a subset of places that are eligible to receive a target meaning.
3. To find the most likely places for the target meaning, apply multi-criteria evaluation (section 6.1) to the eligible places. For personal places, apply it separately to the set of places of each person; for public places, apply to all places together. Select the top scoring places as candidates for receiving the target meaning. Depending on the target meaning, do this in one of the two ways:
   - 3.1 If the evaluation is applied to personal places, and a person typically has a single place with the target meaning (e.g., ‘home’), select the places with the highest scores (note that there may be several such places for a person).
   - 3.2 If the evaluation is applied to public places, or if a person may have several places with the target meaning, choose a score threshold and select the places with the scores not less than the threshold.
4. Examine the temporal distribution of the visits to the best scoring places and check its correspondence to the expected pattern for the target meaning. If the correspondence is not good enough, try to improve it by changing the attribute weights and/or including additional relevant attributes and/or changing the score threshold.
5. If there are land use classes or POI types relevant and/or irrelevant to the target meaning, watch the proportions of land use classes or frequencies of POI types (section 6.2). Test the impact of different attributes and modifications of the attribute weights on these proportions. The goal is to increase the proportions or frequencies of the relevant land use or POI types and decrease those of the irrelevant types. If good results cannot be reached with the current eligibility filter and evaluation criteria, change the filter or include additional criteria.
6. After gaining confidence in the validity of the candidate place selection, assign the target meaning to the selected places (section 6.3).

The following remarks provide further details and show possible modifications to the procedure.

Clustering of time patterns of place visits (section 5) can produce, among others, clusters of time patterns that correspond quite well to certain target meanings. The clustering result can thus be used as a relevant attribute for selecting eligible places for these target meanings. The analyst can select appropriate clusters using a cluster-based filter. Clustering can also reveal temporal patterns that have not been anticipated by the analyst. By observing such a pattern, the analyst may arrive at one of expected target meanings, or find new plausible meanings. The guesses about the plausible meanings can be checked against the LU or POI data.

Filter-based meaning assignment. As described in section 6.3, a meaning can be assigned to a subset of places based on filtering and not on scores or ranks, which means that step 3 of the procedure may be skipped. This is done in the following cases:

1. There are individuals having unique eligible places for the target meaning. Multi-criteria evaluation is not applicable in this case. The eligible places belonging to these individuals can be selected through the related set filter (section 6.4).
2. Existence of multiple places with the target meaning is usual and expected.

Multiple home and work places. While it is not very typical that people have more than one home or work place, such cases do exist. However, it is reasonable to distinguish true cases of multiple home or work places from cases of splitting large places into smaller ones when using the place extraction algorithm (section 3). In the latter case, several places will be close in space, and this can be used as a distinguishing criterion. Based on the above argument, we propose the following way to deal with multiple eligible places for the same target meaning. First, multi-criteria evaluation (step 3) is applied and the topmost-ranking places are selected. After validating the selection in steps 4-5, the target meaning is assigned to these places. Then the distances of the remaining places to the selected places are computed separately for each individual. Eligible places located closely to the selected places (i.e., within a chosen distance threshold, such as for example 1 km) can be treated as parts of the same places and can be assigned the same target meaning. The places that have received meanings are excluded from further consideration through filtering. The remaining eligible places may receive target meaning 'second home' or 'second work' after validation in steps 4-5.

8 ANALYSIS OF SEMANTIC TRAJECTORIES

After assigning semantic labels to personal and public places, the original mobility data (section 2.1) are transformed into semantic trajectories (section 2.3) in the following way. The sequence of records of each individual is divided into daily trajectories taking 4:00 as the beginning hour of a day. From our experiences with many mobility datasets, we know that the total number of recorded activities is usually minimal at this hour; still, another hour can also be chosen. For each trajectory point, the tool tries to find a place containing it. The personal places of the individual are checked first. If no place is found, or a place found has no assigned meaning, the tool checks the public places. If neither personal nor public place is found, the point is skipped (i.e., treated as occasional). If the tool has found a personal or public place with an assigned meaning, the semantic label is attached to the trajectory point. Otherwise, if the point is contained in one of k most frequently visited personal places of the individual (where k is a parameter; in our San Diego experiment, we used k=5), the frequency rank of this place (i.e., 1, 2, …, k) is attached to the point, to allow subsequent checking for relatively frequently visited places that could not be interpreted. Otherwise, the point receives the label "n/a".

After this step, the semantic labels attached to the trajectory points are treated as references to semantic places 'home', 'work', 'shop', etc., located in an abstract semantic space [27]. The set of trajectories can be aggregated by semantic places analogously to aggregation by geographic places [4][5]. Hence, the trajectories are aggregated spatially by the abstract semantic places and temporally by hourly intervals. When the time span covering the data set is not too long, like in the VAST Challenge example, it may be divided into intervals based on the linear time model. Thus, a period of 2 weeks will be divided into 336 (=2×7×24) hourly intervals. A longer time span, as in the San Diego example, can be partitioned based on the cyclic time model.

Specifically, the weekly cycle (7 days) is divided into hourly intervals, resulting in 168 (=7×24) intervals. The result of the aggregation is two sets of time series: hourly visits to semantic places and hourly flows (i.e., aggregated moves) between the places. A move between two places is counted if the time interval between the points in these two places does not exceed one hour. The time series of visits and flows are explored to see whether the daily and weekly temporal patterns are realistic.

To visualize the transformed data, we generate a semantic space [27], i.e., a two-dimensional layout of the set of semantic places. This can be done in various ways. In Fig. 16, the layout has been obtained by applying Sammon's mapping [39] to the set of places based on the strength of the links between them, i.e., the counts of the moves. The layout in Fig. 17 has been produced by the graph visualization software Gephi (http://gephi.org/).

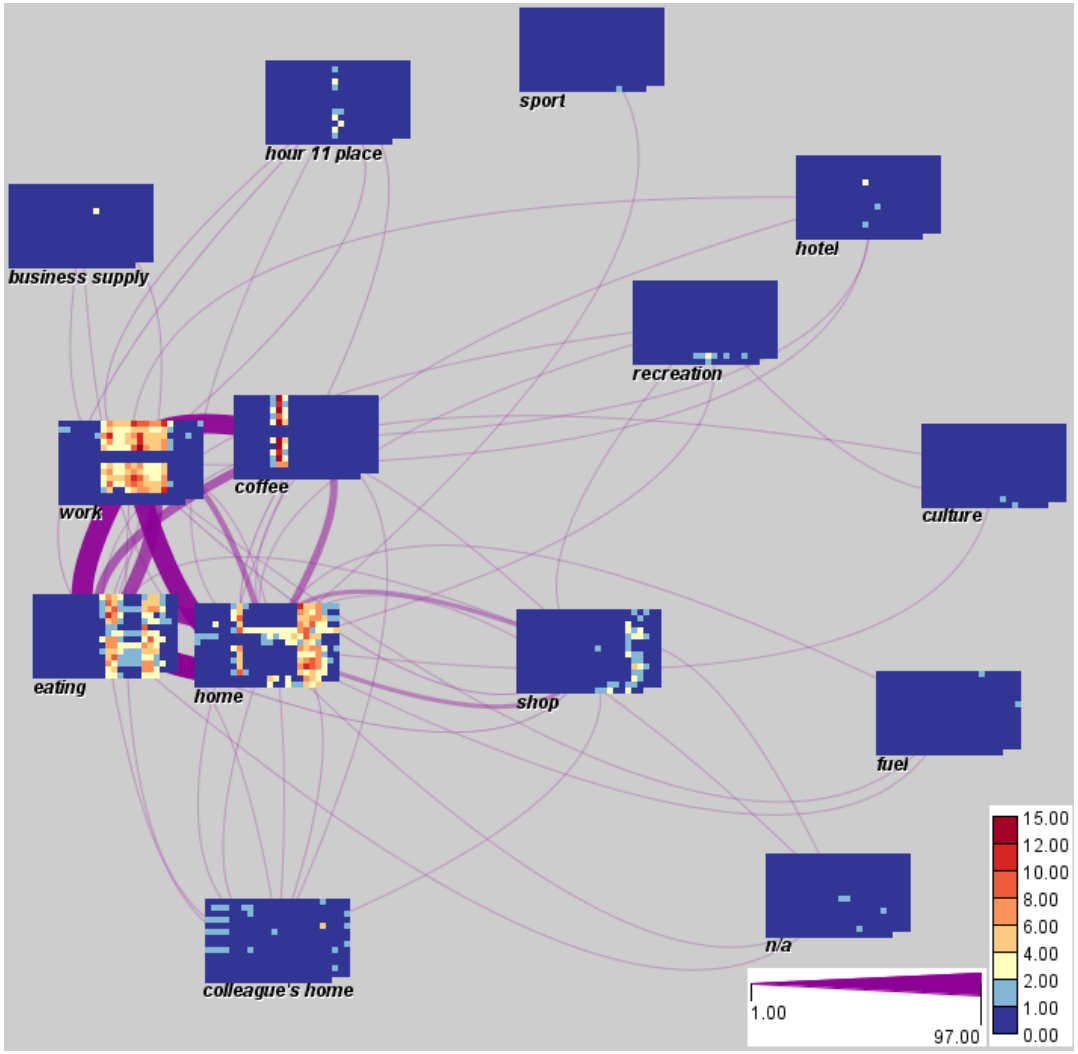

Fig. 16. A map of the semantic space derived from the VAST Challenge data represents the temporal patterns of the visits to the semantic places.

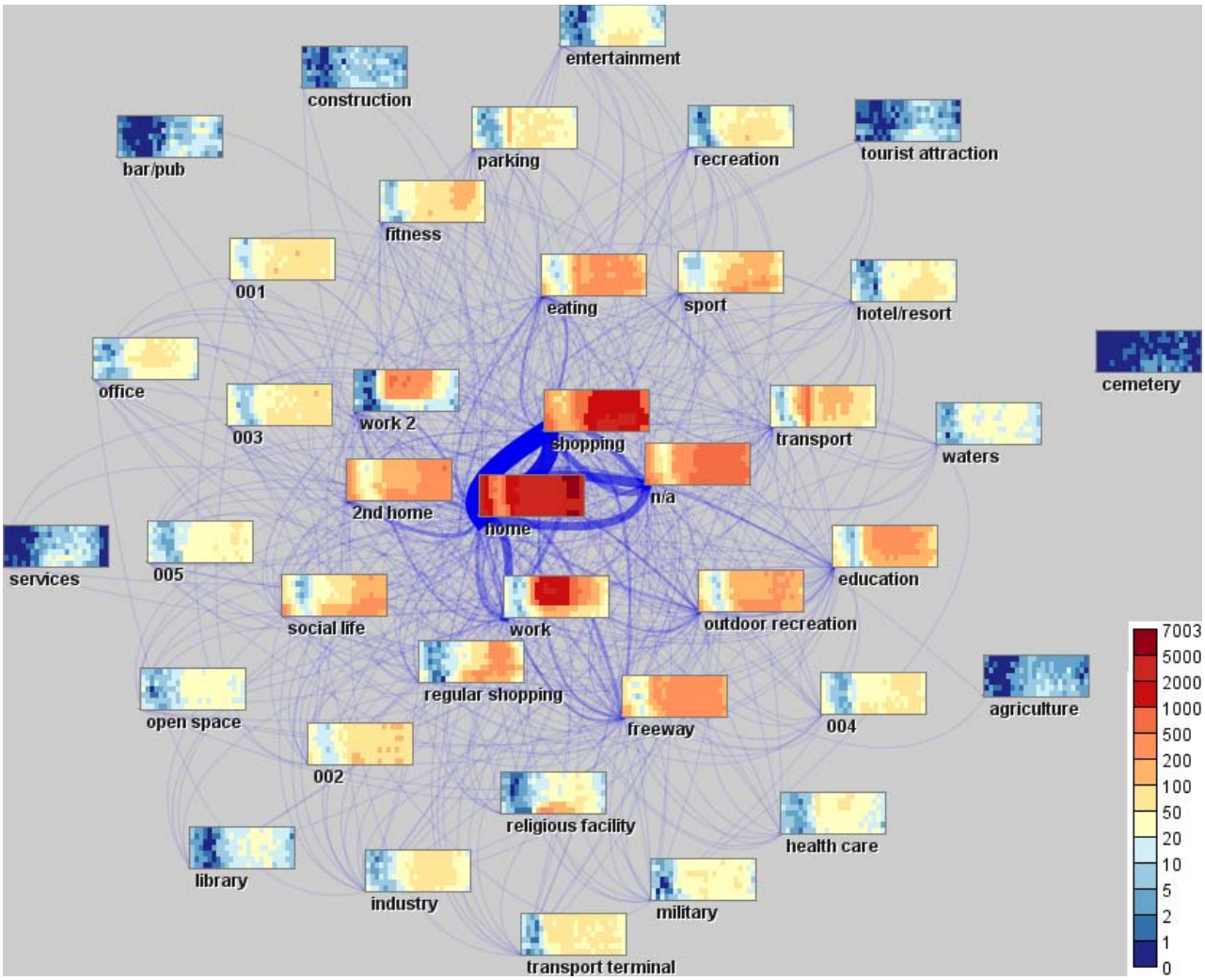

Fig. 17. Semantic information derived from the San Diego data is represented as a semantic space map.

Figs. 16 and 17 show semantic space maps for the VAST Challenge and San Diego examples, respectively. The place visit time series are represented by mosaic diagrams [40]. The rows correspond to 14 consecutive days in Fig. 16 and to 7 days of the weekly cycle in Fig. 17; the columns correspond to 24 hours of the day. The pixels are coloured according to the hourly visit counts. We observe in Fig. 17 that 'shopping' in the San Diego example has the highest attendance after 'home'. This does not necessarily mean that San Diego residents spend more time shopping than doing any other activity. It may just mean that people tweet more often from shopping places than from other types of places. Apart from that, the temporal patterns of the visits to semantically interpreted places observed in both maps correspond to the expected patterns for these types of places or activities.

The flows between the semantic places are shown in Figs. 16 and 17 by curved lines with the curvature increasing towards the destination. The line widths and opacity are proportional to the total number of moves between the origins and destinations. As there is no suitable way to represent flow time series in a single map, and it would be daunting to view and compare

hundreds of maps of the hourly flows, we apply clustering to the hourly flow situations [4][5]. The results of *k*-means clustering with *k*=8 for the VAST Challenge and *k*=7 for San Diego are shown in Figs. 18 and 19, respectively. With the selected values of *k*, we could obtain the simplest yet informative temporal patterns. The mosaic diagrams (at the bottom right and at the top left of the images, respectively) have the same structure as in Figs. 16 and 17, but the pixel colours corresponds to the time clusters of similar flow situations. The distributions of the cluster colours prominently adhere to the daily and weekly time cycles. The multiple maps in both images represent the averaged flow patterns for the clusters. In Fig. 19, we have excluded cluster 1 (blue), in which the flow magnitudes are very low. The line widths and opacity are proportional to the mean hourly move counts.

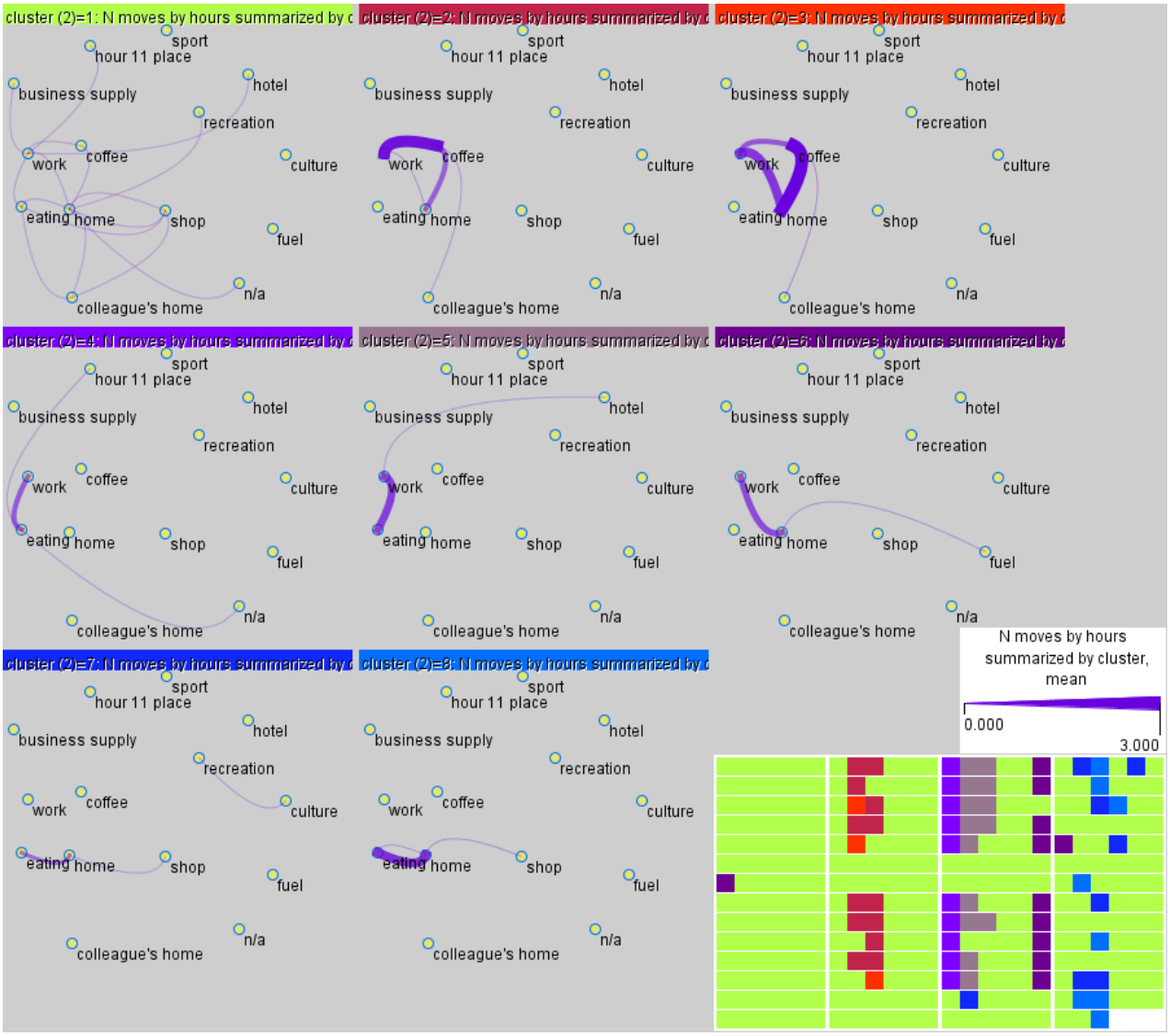

Fig. 18. Averaged flows (summarized movements) between semantic places by time clusters (VAST Challenge).

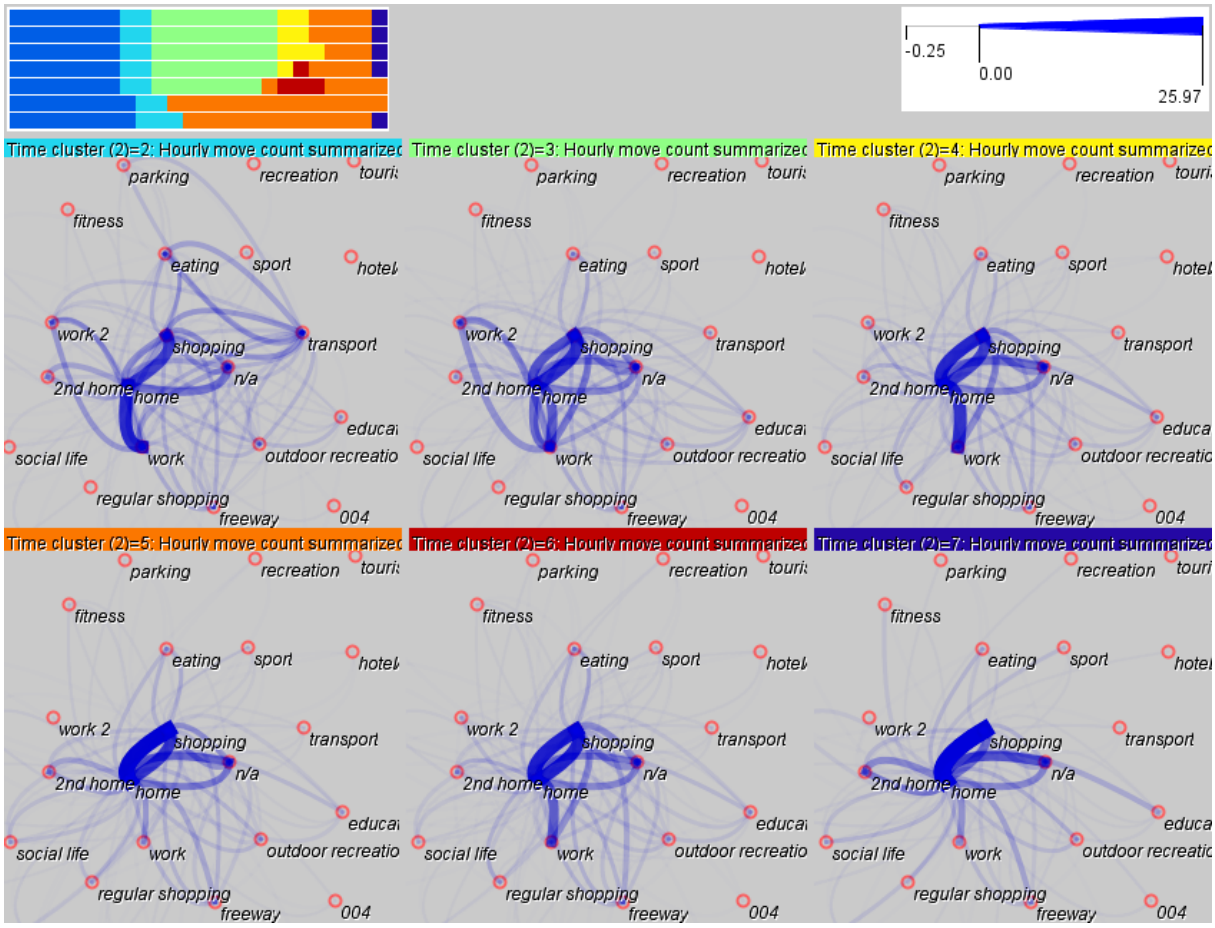

Fig. 19. Flows between semantic places by time clusters (San Diego).

In Fig. 18, we observe quite regular daily routines of the people represented in the VAST Challenge data: during hours 7-8, they usually have coffee and then go to work; at noon, they go for lunch and return back to work in the next one or two hours; then they go home in hour 17 and often go to eat out in the evening. This fully agrees with the available ground truth information.

In Fig. 19, the mean move counts have been transformed to z-scores, i.e., standardized deviations from the clusters' means. This transformation supports disregarding purely numeric differences between the clusters, which are mainly caused by the fact that people tweet less in the morning than in the second half of the day, and reveals differences in the major flow directions. In the morning (cluster 2, cyan), most movements originate from home. Flows to work and to shopping accompanied by the use of transport and parking places are prominent. In the middle of the working days (cluster 3, green), flows from different places to home increase as well as moves between work places ('work' and 'work 2'); the home-work flow decreases as well as the use of transport and parking places. In the afternoon hours (clusters 4 and 6), people mostly return from work to home or go shopping. Cluster 6 (red) occurring on Thursday and Friday differs from cluster 4 (yellow) by smaller flows from work and work 2 to home and bigger flows to places for eating out and social life. In the evenings and on weekends (cluster 5, orange), most of the movements occur between home and shopping but flows to eating establishments and social life are also prominent. Cluster 7 (dark blue) occurring in hour 23 is mostly characterized by returns

to home from different places. All these patterns correspond to our background knowledge about typical human mobility behaviours.

Please note that, although we did not use any information about place visit sequences in our analysis, realistic temporal patterns of movements between places have emerged. This confirms the validity of our methods, in addition to the consistency of the results obtained for the VAST Challenge data with the available ground truth information.

9 Discussion and Conclusion

The problem of deriving significant places from mobility data (i.e., time-stamped geographic positions) that lack semantics is ill-defined; therefore, it cannot be solved algorithmically but instead requires human reasoning. Additional challenges are the necessity to deal with large amounts of data from many individuals and the requirement to respect personal privacy of the individuals. By exploring diverse examples of mobility data, we were able to determine, which visual analytics techniques can support solving the problem at hand. We designed and developed a set of tools meeting the demands of the analytical tasks at hand and tested the effectiveness of these tools in practice, successfully solving the problem for two different datasets. Hence, we can conclude that the problem is solvable and the techniques that we developed work and can be recommended to others. However, due to the ill-defined character of the problem, it is hardly possible to formulate a rigorous procedure. We convey our experience as an informally defined general procedure accompanied by remarks concerning possible particulars and variations (section 7). These guidelines complement our research contribution.

We are aware of the limitations affecting the mobility data that we have used for our research. One of the experiments was done with data from Twitter. It is known that Twitter users are a specific and quite particular group of people, i.e., they cannot be considered as a representative sample of the population. However, this was not a major obstacle for our research, the main goal of which was to develop a methodology for place semantics discovery. Twitter data served as a suitable test dataset, but the methodology can also be applied to other data, in particular, to mobile phone use data, which can represent larger and more diverse section of the population. Second, by example of shopping locations (Figs. 17 and 19), we saw that certain types of places and activities may be over-represented in the data and others may be under-represented. This is a consequence of the episodic character of the original data, where the recorded positions correspond to events that tend to occur more often in some types of places than in others. Mobile phone data can be also affected by such problems. Third, there is a time bias: people tend to tweet more in the afternoons and evenings than in the mornings. Again, the same problem may also exist in mobile phone data. It is necessary to find ways to diminish the place and time biases in analysing population mobility based on episodic movement data. We are conducting further research to address these data-specific issues.

Regarding the location privacy of data containing people's geographic positions, our goal has been to test the possibility of determining place meanings in a privacy-respecting way, i.e., so that analysts only deal with aggregated data and do not access personal data, in particular, individuals' geographic positions. Our research demonstrated that this is practically possible. In two experiments, we used only displays of aggregated data and did not use geographic representations, apart from checking the largest places for choosing right parameters for place extraction (section 3).

To conclude, we have presented a visual analytics approach to the problem of scalable and personal privacy-friendly extraction and semantic interpretation of personal and public places from episodic movement data reflecting human mobility and activities. Our contribution consists of a set of computational and visual techniques, and guidelines for solving the problem with the use of these techniques. We have also proposed methods for evaluation and validation of the results. The approach has been successfully tested on an artificial dataset with known ground truth and on a large real world dataset. The extraction of significant places and their meanings is a necessary step towards reconstruction of mobility diaries, which is the topic of our ongoing research. It can also provide valuable information to researchers of human mobility, who may be interested in studying the variety of individual mobility behaviours. Our research shows that such studies can be done without compromising personal location privacy.

10 REFERENCES

[1] E. Bonabeau. Agent-based modelling: Methods and techniques for simulating human systems. *PNAS Proceedings of the National Academy of Sciences of the USA*, 99(Suppl. 3): 7280-7287, May 2002.

[2] G. Andrienko, N. Andrienko, and S. Wrobel. Visual Analytics Tools for Analysis of Movement Data. *ACM SIGKDD Explorations*, 9(2): 38-46, Dec. 2007.

[3] J. Bertin. *Semiology of Graphics. Diagrams, Networks, Maps*. University of Wisconsin Press, Madison, 1983.

[4] G. Andrienko, N. Andrienko, H. Stange, T. Liebig, and D. Hecker. Visual Analytics for Understanding Spatial Situations from Episodic Movement Data. *Künstliche Intelligenz*, 26(3): 241-251, 2012.

[5] G. Andrienko, N. Andrienko, P. Bak, D. Keim, and S. Wrobel. *Visual Analytics of Movement*. Springer, 2013.

[6] G. Andrienko, N. Andrienko, G. Fuchs, A.-M. Olteanu-Raimond, J. Symanzik, and C. Ziemlicki. Extracting semantics of individual places from movement data by analyzing temporal patterns of visits. *Proc. ACM SIGSPATIAL Workshop Computational Models of Place*, 2013.

[7] A.T. Palma, V. Bogorny, B. Kuijpers, and L.O. Alvares. A clustering-based approach for discovering interesting places in trajectories. In *Proc. 23rd Annual ACM Symposium on Applied Computing* (*ACM SAC)*, New York, USA: ACM Press, pp. 863-868, 2008.

[8] M. Zimmermann, T. Kirste, and M. Spiliopoulou. Finding Stops in Error-Prone Trajectories of Moving Objects with Time-Based Clustering. *In Intelligent Interactive Assistance and Mobile Multimedia Computing*, pp.275-286, 2009.

[9] J. Rocha, G. Oliveira, L. Alvares, V. Bogorny, and V. Times. Db-smot: A direction-based spatio-temporal clustering method. In *Proc. 5th IEEE Int. Conf. Intelligent Systems*. University of Westminster, London, UK, pp. 114-119, 2010.

[10] C. Parent, S. Spaccapietra, C. Renso et al. Semantic Trajectories Modeling and Analysis. *ACM Computing Surveys*, 45(4), article 42, 2013.

[11] S. Phithakkitnukoon, T. Horanont, G. Di Lorenzo, R. Shibasaki, and C. Ratti. Activity-Aware Map: Identifying human daily activity pattern using mobile phone data. In *Proc. Int. Conf. Pattern Recognition, Workshop on Human Behavior Understanding*, Springer, Heidelberg, pp. 14-25, 2010.

[12] J.L. Toole, M. Ulm, M.C. González, and D. Bauer. Inferring land use from mobile phone activity. In Proc. ACM SIGKDD Int. Workshop Urban Computing (UrbComp'12), pp. 1-8, 2012.
[13] R. Ahas, S. Silm, O. Järv, E. Saluveer, and M. Tiru. Using Mobile Positioning Data to Model Locations Meaningful to Users of Mobile Phones. *Journal of Urban Technology*, 17(1), pp.3-27, April 2010.
[14] 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*, pp: 133–151, 2011.
[15] C.M. Schneider, V. Belik, T. Couronné, Z. Smoreda, and M.C. González. Unraveling Daily Human Mobility Motifs. *Journal of the Royal Society Interface*, 10 20130246, 2013.
[16] C.M. Schneider, C. Rudloff, D. Bauer, and M.C. González. Daily travel behavior: Lessons from a week-long survey for the extraction of human mobility motifs related information. In *Proc. 2nd ACM SIGKDD Int. Workshop Urban Computing (UrbComp'13)*, Article No. 3, 2013.
[17] Z. Cheng, J. Caverlee, K.Y. Kamath, and K. Lee. Toward Traffic-Driven Location-Based Web Search. In *Proc. 20th ACM Conf. Information and Knowledge Management (CIKM 2011)*, 2011.
[18] M. Ye, K. Janowicz, W.-C. Lee, and C. Mülligann. What you are is When you are: The Temporal Dimension of Feature Types in Location-based Social Networks. In *Proc. 19th ACM SIGSPATIAL Int. Conf. Advances in Geographic Information Systems*, 2011.
[19] H. Liu, B. Luo, and D. Lee. Location Type Classification Using Tweet Content. In *Proc. 11th Int. Conf. Machine Learning and Applications (ICMLA 2012)*, Vol. 1, pp. 232-237, 2012.
[20] VAST challenge 2011. http://hcil.cs.umd.edu/localphp/hcil/vast11/
[21] H. Bosch, D. Thom, F. Heimerl, E. Püttmann, S. Koch, R. Krüger, M. Wörner, and T. Ertl. ScatterBlogs2: Real-Time Monitoring of Microblog Messages Through User-Guided Filtering. In *IEEE Conference on Visual Analytics Science and Technology (VAST)*, 2013.
[22] R. Krüger, D. Thom, and T. Ertl. Visual Analysis of Movement Behavior using Web Data for Context Enrichment. In: *IEEE Pacific Visualization Symposium (PacificVis)*, 2014.
[23] J. Cuellar, M. Ochoa, and R. Rios, Indistinguishable Regions in Geographic Privacy. In *Proc. 27th Annual ACM Symposium Applied Computing (SAC 2012)*, S. Ossowski and P. Lecca, Eds., ACM, pp. 1463-1469, 26-30 March 2012.
[24] G. Andrienko, N. Andrienko, D. Keim, A. MacEachren, and S. Wrobel. Challenging Problems of Geospatial Visual Analytics. *Journal of Visual Languages and Computing*, 22(4): 251-256, 2011.
[25] G. Andrienko and N. Andrienko. Privacy issues in geospatial visual analytics. *Advances in Location-Based Services*, 239–246, 2012.
[26] Monreale, G. Andrienko, N. Andrienko, F. Giannotti, D. Pedreschi, S. Rinzivillo, and S. Wrobel. Movement Data Anonymity through Generalization. *Transactions on Data Privacy*, 3(3): 91-121, 2010.
[27] N. Andrienko, G. Andrienko, and G. Fuchs. Towards Privacy-Preserving Semantic Mobility Analysis. In *Proc. EuroVA, Int. Workshop on Visual Analytics 2013*, EuroGraphics, 2013.
[28] Dasgupta and R. Kosara. Adaptive Privacy-Preserving Visualization Using Parallel Coordinates. *IEEE Trans. Visualization and Computer Graphics*, 17(12): 2241-2248, 2011.

[29] VAST Challenge 2014: Mini-Challenge 2. http://www.vacommunity.org/VAST+Challenge+2014%3A+Mini-Challenge+2

[30] G. Andrienko, N. Andrienko, H. Bosch, T. Ertl, G. Fuchs, P. Jankowski, and D. Thom. Discovering Thematic Patterns in Geo-Referenced Tweets through Space-Time Visual Analytics. *Computing in Science and Engineering*, 15(3): 72-82, 2013.

[31] G. Andrienko, N. Andrienko, C. Hurter, S. Rinzivillo, and S. Wrobel. Scalable Analysis of Movement Data for Extracting and Exploring Significant Places. *IEEE Trans. Visualization and Computer Graphics*, 19(7): 1078-1094, 2013.

[32] N. Andrienko and G. Andrienko. Spatial generalization and aggregation of massive movement data. *IEEE Trans. Visualization and Computer Graphics*, 17(2): 205-219, 2011.

[33] Y. Tsuda, Q. Kong, and T. Maekawa. Detecting and correcting WiFi positioning errors. In *Proc. ACM Int. Joint Conf. Pervasive and Ubiquitous Computing (UbiComp '13)*. ACM, New York, NY, USA, 777-786, 2013.

[34] P.A. Zandbergen. Accuracy of iPhone Locations: A Comparison of Assisted GPS, WiFi and Cellular Positioning. *Transactions in GIS*, 13(s1): 5–26, 2009.

[35] S. Greco (ed.). *Multiple Criteria Decision Analysis: State of the Art Surveys*. Springer, Berlin, 2005.

[36] J. Malczewski. *GIS and Multicriteria Decision Analysis*. New York: John Wiley & Sons, 1999.

[37] P. Jankowski, N. Andrienko, and G. Andrienko. Map-Centered Exploratory Approach to Multiple Criteria Spatial Decision Making. International Journal Geographical Information Science, 15(2): 101-127, 2001.

[38] N. Andrienko and G. Andrienko. Informed Spatial Decisions through Coordinated Views. *Information Visualization*, 2 (4): 270-285, 2003.

[39] J. W. Sammon. A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, 18: 401–409, 1969.

[40] G. Andrienko and N. Andrienko. Spatio-temporal aggregation for visual analysis of movements. In *Proc. IEEE Visual Analytics Science and Technology (VAST 2008)*, IEEE Computer Society Press, pp.205-206, 2008.

Appendices to paper

Scalable Interactive Discovery of Place Semantics from Human Mobility Traces

Appendix I. Preparation of data from VAST Challenge 2014 Mini-Challenge 2

The original dataset provided for the VAST Challenge 2014 Mini-Challenge 2 [1] consists of simulated tracks of cars with duration of two weeks. The records include timestamps, car identifiers, and coordinates. We used the tracks of 35 personal cars and ignored the tracks of 5 remaining vehicles utilized only for business purposes. The full tracks cannot serve as an example of episodic mobility data because of a high temporal resolution (1 second), which allows determining the exact times of coming to each visited place and leaving it.

To have a suitable example of episodic data, we extracted a subset of the car position records in the following way. First, we extracted the events of stopping for at least one minute. Stops are reflected in the data as temporal gaps between consecutive position records, since, according to the description of the data, the positions were recorded only when the vehicles were moving. For each stop event, we took both the last record before the gap (i.e., stop start) and the first record after the gap (i.e., stop end). Additionally, we extracted the first and last records of each track. We obtained in total 6,068 position records, which is less than 1% of the original 613,077 records. From these 6,068 records, we extracted a 25% random sample (1,469 records). It imitates the properties of episodic mobility data, where a stop at a location may be represented by one or more records, or it may not be represented at all.

No data similar to land use or POI data were provided for the challenge. As the underlying territory for the car tacks is fictitious, existing databases or map feature services cannot give us suitable information about places. To create a substitute for POI data, we utilized simulated credit card transaction records, also available for the challenge. Each record includes a timestamp, the name of the location where the card was used for payment, the amount paid, and the first and last names of the customer. We complemented these records with the identifiers of the cars used by the customers and the types (semantic categories) of the locations, which include ‘eating’, ‘coffee’, ‘shop’, ‘hotel’, ‘sport’, ‘culture’, and ‘business supply’. A fragment of the card transaction data table with the added attributes is shown in Fig.1. The column “Interpretation” contains the semantic categories of the locations, and the column “CARID” contains the identifiers of the cars.

	LOCATION	Interpretation	TIMESTAMP	PRICE	CARID	LASTNAME	FIRSTNAME
47	Abila Zacharo	eating	09/01/2014 13:41:00	89.41	1	Alcazar	Lucas
89	Abila Zacharo	eating	18/01/2014 13:32:00	16.59	1	Alcazar	Lucas
98	Albert"s Fine Clothing	shop	06/01/2014 20:26:00	276.9	1	Alcazar	Lucas
384	Frydos Autosupply n" M	shop	13/01/2014 19:20:00	10000	1	Alcazar	Lucas
406	Gelatogalore	eating	07/01/2014 13:37:00	21.52	1	Alcazar	Lucas
550	Guy"s Gyros	eating	16/01/2014 13:28:00	10.27	1	Alcazar	Lucas
594	Hallowed Grounds	coffee	06/01/2014 07:55:00	8.05	1	Alcazar	Lucas
598	Hallowed Grounds	coffee	07/01/2014 07:46:00	84.44	1	Alcazar	Lucas
609	Hallowed Grounds	coffee	08/01/2014 07:56:00	12.86	1	Alcazar	Lucas
612	Hallowed Grounds	coffee	09/01/2014 07:50:00	34.45	1	Alcazar	Lucas
646	Hallowed Grounds	coffee	16/01/2014 08:05:00	12.19	1	Alcazar	Lucas
653	Hallowed Grounds	coffee	17/01/2014 08:04:00	9.4	1	Alcazar	Lucas
655	Hippokampos	eating	06/01/2014 13:21:00	28.23	1	Alcazar	Lucas
679	Hippokampos	eating	08/01/2014 13:43:00	39.8	1	Alcazar	Lucas
708	Hippokampos	eating	11/01/2014 13:37:00	75.62	1	Alcazar	Lucas
720	Hippokampos	eating	12/01/2014 14:06:00	71.99	1	Alcazar	Lucas
731	Hippokampos	eating	13/01/2014 13:28:00	30.51	1	Alcazar	Lucas
1034	Kronos Mart	shop	19/01/2014 03:45:00	194.51	1	Alcazar	Lucas
1086	Ouzeri Elian	eating	08/01/2014 21:16:00	30.81	1	Alcazar	Lucas
1094	Ouzeri Elian	eating	10/01/2014 13:16:00	30.71	1	Alcazar	Lucas

Fig. 1. A fragment of the table with the credit card transaction data enriched with location interpretations.

	X	Y	time	Movement event type	Identifier of nearest transaction event within -15..90 minutes	Time distance to the nearest transaction event; minutes	Transaction event time	Transaction location	Location type
1	24.885931	36.063713	01/06/2014 07:57:01	stop end	594	2.02	06/01/2014 07:55:00	Hallowed Grounds	coffee
1	24.879574	36.048027	01/06/2014 08:04:09	stop start					
1	24.882586	36.0665	01/06/2014 19:36:01	stop end					
1	24.85632	36.075283	01/06/2014 19:49:01	stop start	98	-36.98	06/01/2014 20:26:00	Albert"s Fine Clothing	shop
1	24.879574	36.048023	01/06/2014 23:01:01	stop start					
1	24.879568	36.048115	01/07/2014 01:10:01	stop end					
1	24.870821	36.051968	01/07/2014 18:55:52	stop start					
1	24.882656	36.066475	01/07/2014 20:52:01	stop end					
1	24.885891	36.063663	01/08/2014 07:49:04	stop start	609	-6.93	08/01/2014 07:56:00	Hallowed Grounds	coffee
1	24.87957	36.048115	01/08/2014 17:51:01	stop end					
1	24.882559	36.066498	01/08/2014 19:18:01	stop end					
1	24.879568	36.04803	01/08/2014 21:29:01	stop end	1086	13.02	08/01/2014 21:16:00	Ouzeri Elian	eating
1	24.882612	36.066456	01/09/2014 03:20:01	stop end					
1	24.885885	36.06365	01/09/2014 07:23:04	stop start	612	-26.93	09/01/2014 07:50:00	Hallowed Grounds	coffee
1	24.879574	36.048023	01/09/2014 12:09:01	stop end					
1	24.851015	36.063366	01/09/2014 13:45:01	stop end	47	4.02	09/01/2014 13:41:00	Abila Zacharo	eating
1	24.885914	36.063663	01/10/2014 08:08:01	stop end					
1	24.87078	36.051926	01/10/2014 12:23:05	stop start	1094	-52.92	10/01/2014 13:16:00	Ouzeri Elian	eating
1	24.879566	36.04803	01/10/2014 13:23:05	stop start					
1	24.860416	36.085472	01/10/2014 19:20:01	stop start					
1	24.882536	36.06642	01/11/2014 18:42:01	stop end					
1	24.882618	36.066463	01/11/2014 19:52:07	stop start					
1	24.857595	36.076656	01/12/2014 14:11:01	stop end	720	5.02	12/01/2014 14:06:00	Hippokampos	eating

Fig. 2. Stop records enriched with information about the nearest in time transaction events.

The card transaction data as such cannot substitute POI data because there are no coordinates of the locations. We solved this problem by linking transaction records to car stop records based on the times of the transactions and the stops. For each stop record, we selected the closest in time transaction record with the same car identifier as in this stop record. This was done differently for stop starts and stop ends. For stop starts, the search for the closest card transaction was done forward in time within the interval of 1.5 hours, assuming that customers usually pay after spending some time at a location. For stop ends, the search was done backward in time within the interval of 15 minutes, assuming that customers usually pay shortly before leaving a location.

Not all car transaction records turned to be suitable. For three coffee shops, the transaction timestamps were not trustable, since the time of the day in all of them was 12:00:00. These records were not used.

We were able to find the closest transactions for 1,849 out of 6,068 stop records (30.5%). The location types of the closest transactions were attached to the stop records; a fragment of the table with the resulting data is shown in Fig. 2. These assignments need to be used with caution. Since the people did not pay by credit cards during all of their stops, some stops might be

associated with transactions made elsewhere. Still, for the places where people were supposed to pay, it can be expected that the majority of the stop records have got right assignments of location types. Of course, this does not apply to the three coffee shops with uniform transaction times. The stops at these coffee shops could get either no location types or wrong location types from irrelevant transaction records. In the following, we shall show how these data can be used with taking into account the possible errors.

We would like to stress that, although the conditions of the challenge did not require it, we analysed the data in a privacy-respecting way, i.e., without looking at any personal data.

REFERENCES

[1] VAST Challenge 2014: Mini-Challenge 2. http://www.vacommunity.org/VAST+Challenge+2014%3A+Mini-Challenge+2

Appendices to paper

Scalable Interactive Discovery of Place Semantics from Human Mobility Traces

Appendix II. INTERACTIVE TOOLS FOR DERIVATION OF PLACE ATTRIBUTES

For each place extracted from mobility data, an automated tool derives a two-dimensional time series of the place visits by hours of the day for different days of the week, 168 (=24×7) counts in total. For personal places, only the visits of the place owners are counted. Counts of visits are not the same as counts of points. If two consecutive points of a person fit in the same place and the same hour, they are treated as representing the same visit.

We have developed an interactive tool for convenient derivation of further attributes from the two-dimensional time series of the place visit counts. Thus, it may be necessary to compute the number or percentage of place visits that fit in the work time, i.e., in the hours from 05 to 18 during work days. The UI of the tool is shown in Fig. 3. To select the hourly intervals that need to be summed, the user clicks on the corresponding cells, columns, or rows of the matrix. The rows correspond to days of the week and the columns to hours of the day. The sums may be normalized as ratios or percentages of the user-chosen attribute, e.g., the total visit count.

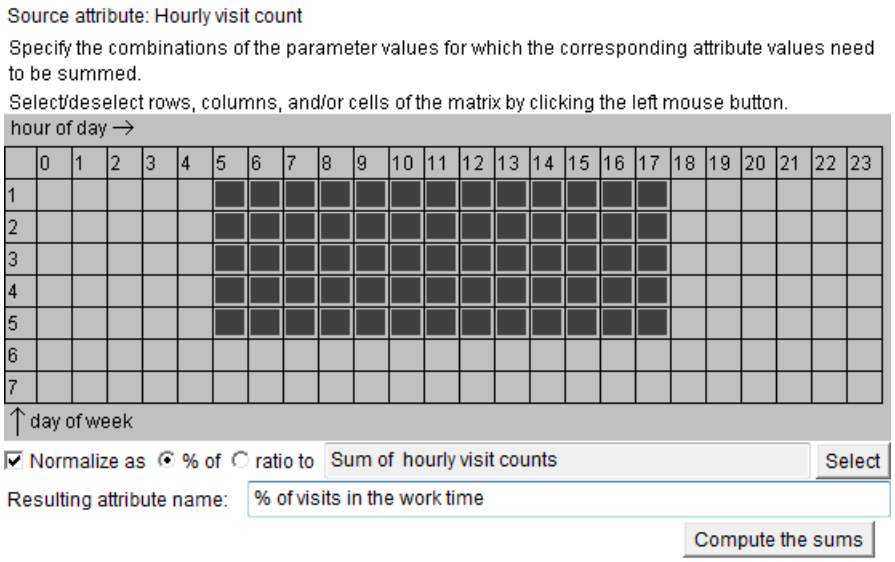

Fig. 3. Selecting elements of 2-dimensional time series for summing.

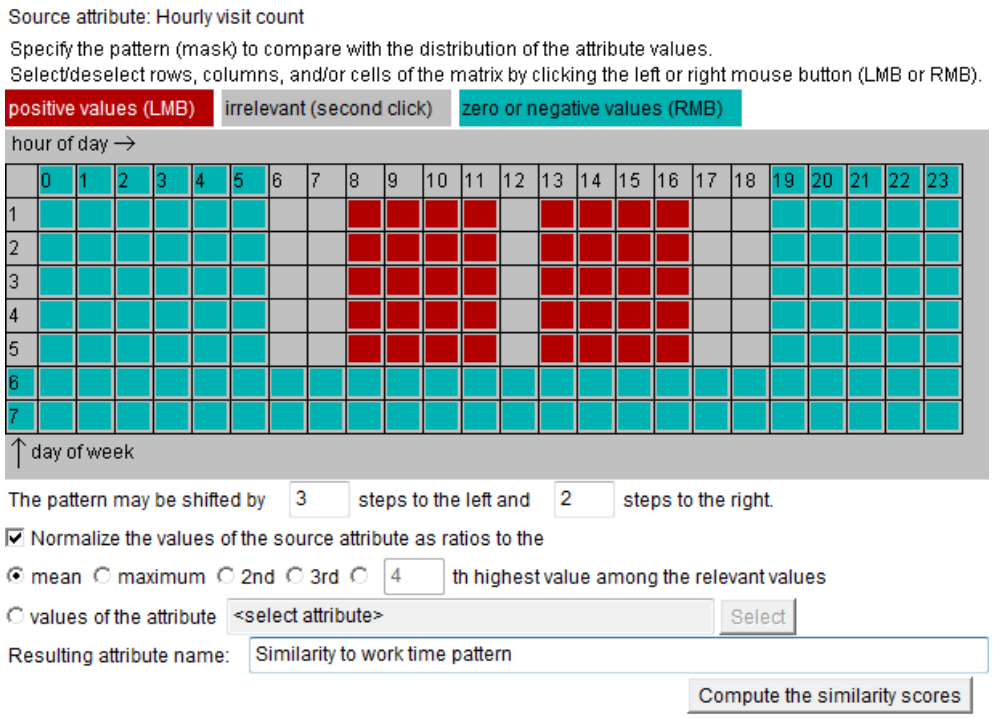

Fig. 4. Interactive specification of a 2d temporal pattern for computing similarity scores.

A similar interface (Fig. 4) has been built for computing the degrees of similarity in temporal patterns of place visits to an arbitrary, user-defined pattern. The user "paints" the matrix cells in three colours. The red colour means that the corresponding component of the time series has a positive impact on the similarity score, i.e., its value will increase the score. The cyan colour means that the component has a negative impact, i.e., its value will diminish the score. The grey colour is neutral, i.e., the corresponding component has no impact. Fig. 4 shows an example of a painted matrix for a work time pattern. According to this pattern, a person is expected to be present at a place from 8:00 until 17:00 in the work days, possibly, with a lunch break in between, and is not expected to be present before 6:00, after 19:00, and on the weekend. Of course, different people may start and finish their work at different times. To account for such differences, the pattern may be shifted to the left and/or to the right by the user-specified number of hours. In Fig. 4, the user allows the tool to shift the pattern by up to 3 hours to the left and up to 2 hours to the right, thus covering the work time intervals in the range from 5-14 to 10-19. The tool computes the similarity scores for all possible positions of the pattern and selects the maximal score. The original values involved in the computation may be normalized; the possible normalization options can be seen in Fig. 4. The resulting scores are scaled to the range from -1 (completely opposite) to 1 (perfectly matching).

We have also developed a thematic enrichment tool that derives various aggregate attributes of places from user-chosen attributes of the points belonging to these places. For each place, the tool selects from the database the points contained in this place. For personal places, only the points of the place owners are selected. The aggregate attributes that can be derived depend on the types of the original attributes:

- Numeric: minimum, maximum, sum, mean, standard deviation, and arbitrary percentiles.

- Qualitative: (Q1) the number of distinct categories; (Q2) *k* most frequent categories (i.e., having ranks 1, 2, …, *k* in the descending frequency order; *k* is chosen by the user) and their frequencies.
- Textual: (T1) *k* most frequent words and their frequencies. The user can supply a list of stop words to be ignored; (T2) frequencies of occurrences of terms from a user-supplied dictionary. The dictionary may be composed of main terms and their synonyms or related words. Occurrences of related words are counted as occurrences of the main terms.

Land use classes can be attached to places by deriving Q2 from the land use classes of the points. Multiple points contained in the same place may have different land use classes. It may be insufficient to take only one most frequent class. In our San Diego example, we chose *k*=5 to retrieve 5 most frequent land use classes per place.

For Twitter data, which include texts of the posted messages, it is possible to obtain T2, i.e., counts of occurrences of different topics (subjects) people tweeted about, such as "family", "home", "work", "education", "friends", "food", etc. **Error! Reference source not found.**. These counts can be used additionally to land use or POI data; however, in this paper, we do not focus on using Twitter-specific information.

From POI data, counts of different types of POIs inside the places can also be derived as T2. For this purpose, the possible POI types need to be listed as terms in a dictionary.

REFERENCES

[1] G. Andrienko, N. Andrienko, H. Bosch, T. Ertl, G. Fuchs, P. Jankowski, and D. Thom. Discovering Thematic Patterns in Geo-Referenced Tweets through Space-Time Visual Analytics. *Computing in Science and Engineering*, 15(3): 72-82, 2013.

Appendices to paper

Scalable Interactive Discovery of Place Semantics from Human Mobility Traces

Appendix III. ANALYSIS EXAMPLE

We shall demonstrate the use of the proposed tools for place meaning discovery on the example of the dataset constructed from the VAST Challenge data. It is more suitable for demonstration purposes as it is smaller and simpler than the San Diego data; besides, some ground truth is available for it. The analysis of the San Diego data included more steps and would be tedious to describe and to read.

III.1 Analysis of personal places

III.1.1 Identifying home places

We start the analysis of the VAST Challenge data with an attempt to find the home places of the 35 individuals among the 202 personal places we have extracted earlier. We shall describe the process of identifying and labelling home places in much detail, to show how the analysis is done and how the tools are used.

Using the interactive tool shown in Fig. 3, we derive attributes: "% of visits in home time (hours 18-08 + weekend)" and "% of visits in work time (hours 07-19 on week days)" from the hourly counts of place visits. We apply the place ranking tool using these two attributes and attribute "number of different visit days" (computed automatically by the place extraction tool) as criteria (Fig. 5). The attribute "% of visits in work time" is minimized, and the two others are maximized. When all criteria have equal weights, 36 places of 35 distinct owners receive the topmost ranks. After a small increase of the weight of the attribute "% of visits in home time", the number of the topmost ranked places decreases to 35, so that there is a single candidate home place for each individual.

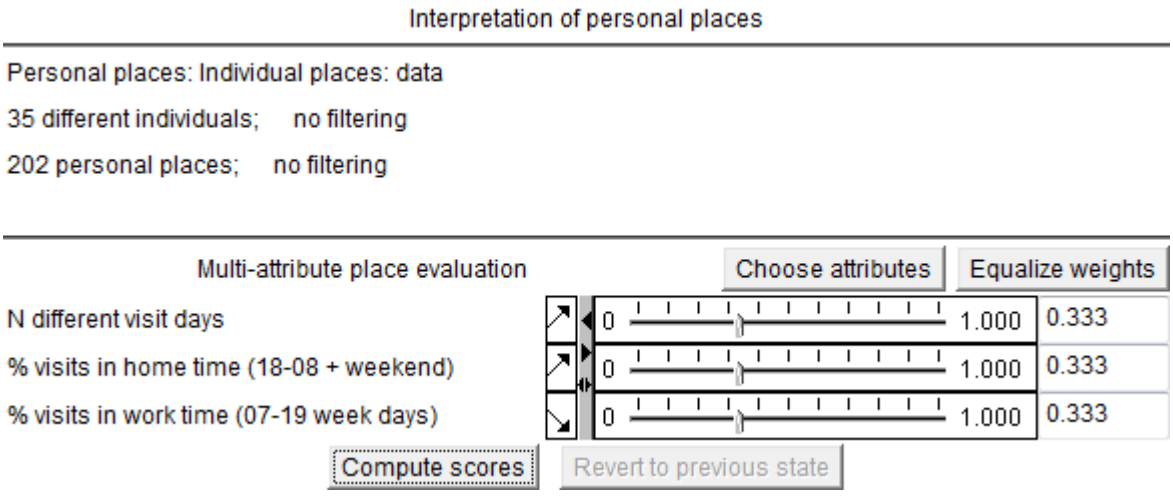

Fig. 5. The multi-criteria ranking tool is used for finding the most likely home places.

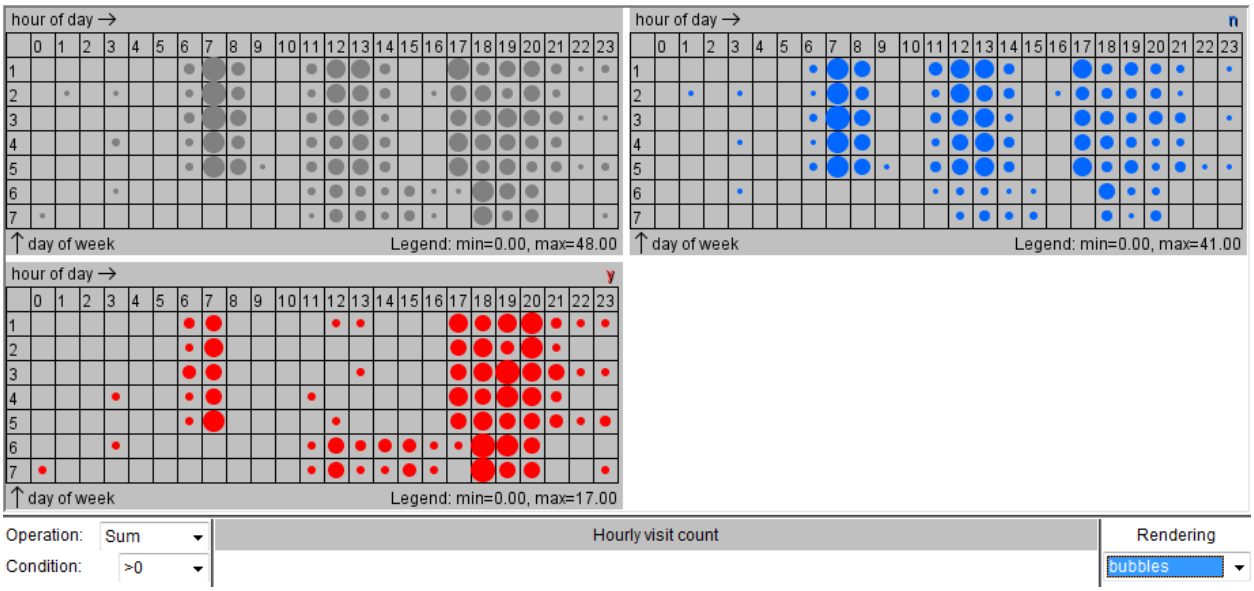

Fig. 6. The results of place ranking for the target meaning 'home' are represented on a 2d time histogram display.

We propagate the place classes ('y' for the topmost ranked places and 'n' for the remaining places) to a 2d time histogram display (Fig. 6). The class 'y' is represented by red colour and the class 'n' by blue colour. We look at the temporal distribution of the stops in the subset of the top ranked places (red) and see that there are some stops at the lunch hours of the week days, which hints that the subset may include eating places. We check this hypothesis using two multi-attribute bar chart displays of the POI types associated with the places. One display summarizes the counts of the stop points labelled by different POI types (Fig. 7 top) and the other display summarizes the percentages of the stops labelled by different POI types (Fig. 7 bottom).

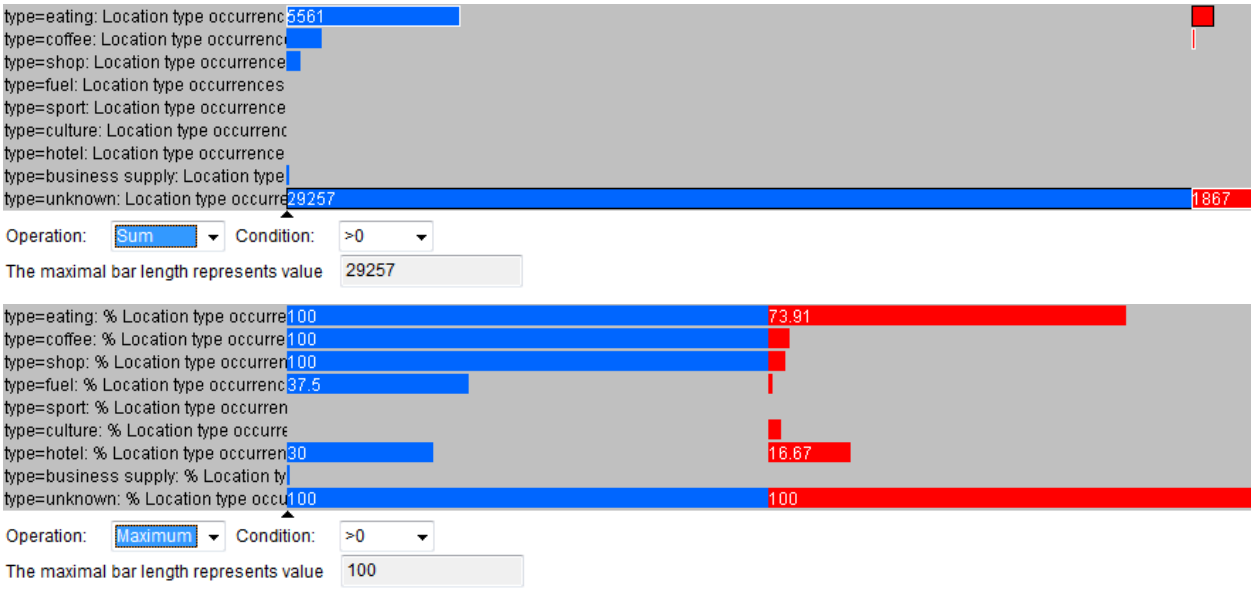

Fig. 7. The multi-attribute bar charts represent the sums of the counts of different POI types (top) and the maximal percentages of the different POI types (bottom) in two classes of places.

The multi-attribute bar chart representing the percentages of the different POI types shows a very high maximum (73.8%) for the POI type ‘eating’, thus confirming the guess.

We try to improve the place selection by changing the weights of the currently used criteria, but this does not help; thence, we need to involve an additional criterion. To lower the ranks of the eating places, which are visited at the lunch time, we compute and employ a new criterion, ‘% of visits in lunch time (hours 12-15) on week days’, which needs to be minimized (Fig. 8). A good result is obtained when the new criterion is given a high weight (0.65), which removes the places visited at the lunch time from the top ranked places.

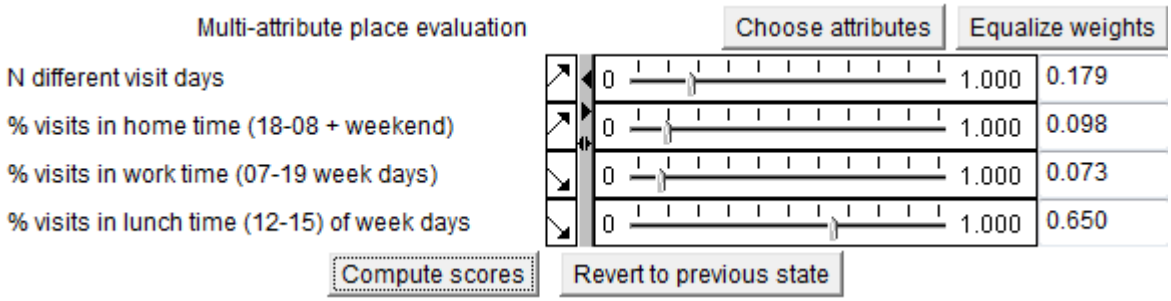

Fig. 8. A new criterion “% visits in lunch time (12-15) of week days” has been added for a better separation of home places from eating places.

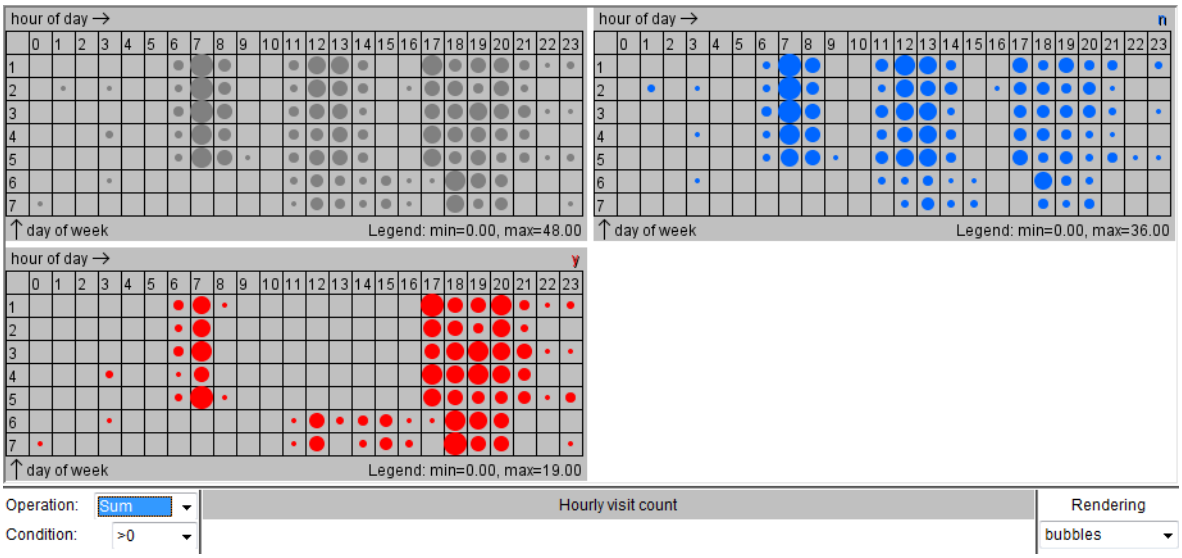

Fig. 9. Improved results of place ranking for the target meaning ‘home’ are represented on a 2d time histogram display.

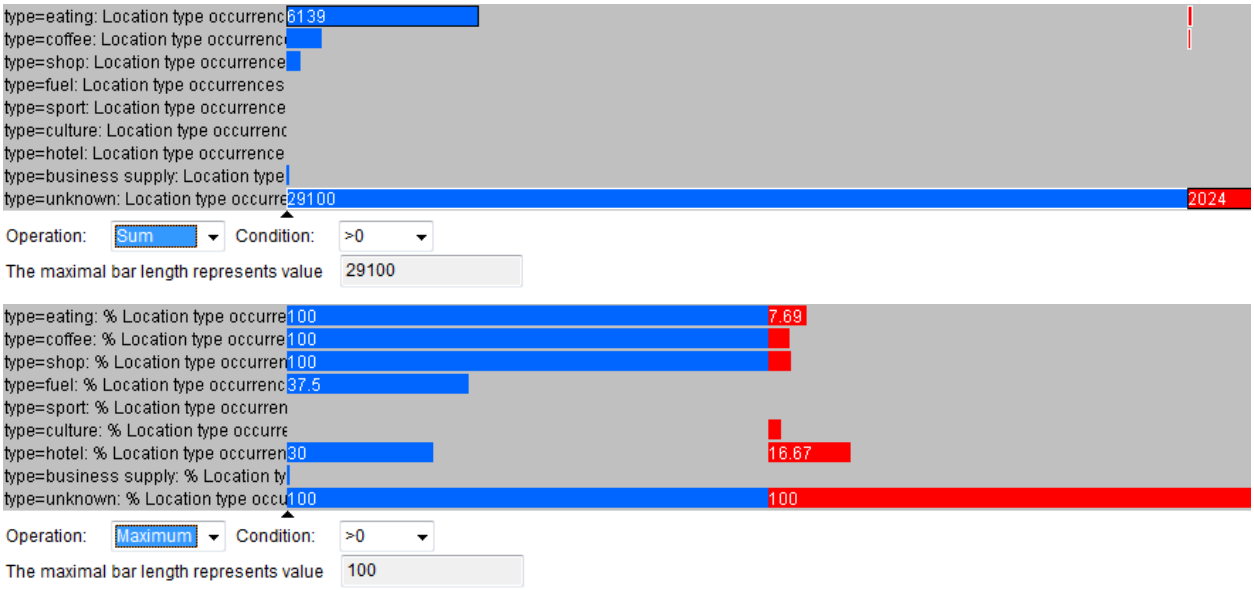

Fig. 10. The multi-attribute bar charts of the POI types confirm that the place ranking for the target meaning ‘home’ has improved after adding a new criterion.

Fig. 9 shows the resulting temporal distributions of the visits in the top ranked places (red) and the remaining places (blue), and Fig. 10 shows the cumulative counts and the maximal percentages of the different POI types for the top ranked places and for the remaining places. The maximal percentages are now only 7.69% for ‘eating’ and even lower for the other POI types, except for ‘hotel’ (16.67%). We filter out the places with high percentages of the POI type ‘hotel’ and re-compute the scores and ranks for the remaining places in a hope to find better candidates for the meaning ‘home’. However, only 34 places of 34 owners could this time receive the best scores. Evidently, one person had no home within the area and stayed in a hotel, which played the role of this person’s home. Based on this reasoning, we cancel the filter and revert the ranking to the previous state. Finally, we assign the meaning ‘home’ to the 35 top ranked places of 35 individuals.

III.1.2 Identifying work places

By filtering, we exclude the places that have already got semantic labels (i.e., the home places) from the further consideration and start the process of identifying work places. We again use the criteria “number of different visit days”, “% of visits in work time” and “% of visits in home time”. The first two are maximized and the third one is minimized. With equal weights, we get 35 candidate work places of 34 distinct persons, i.e., one person has two candidate work places with equal scores.

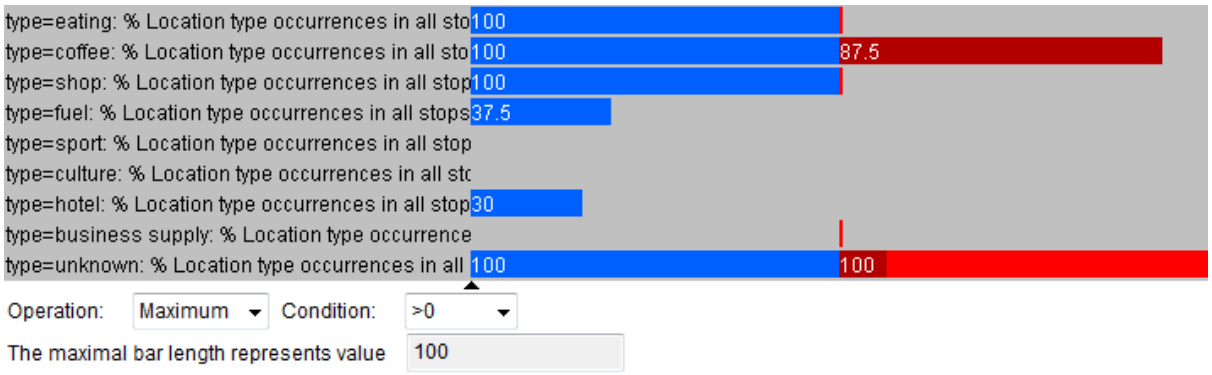

Fig. 11. The multi-attribute bar chart of the POI types reflects the result of the place ranking for the target meaning ‘work’.

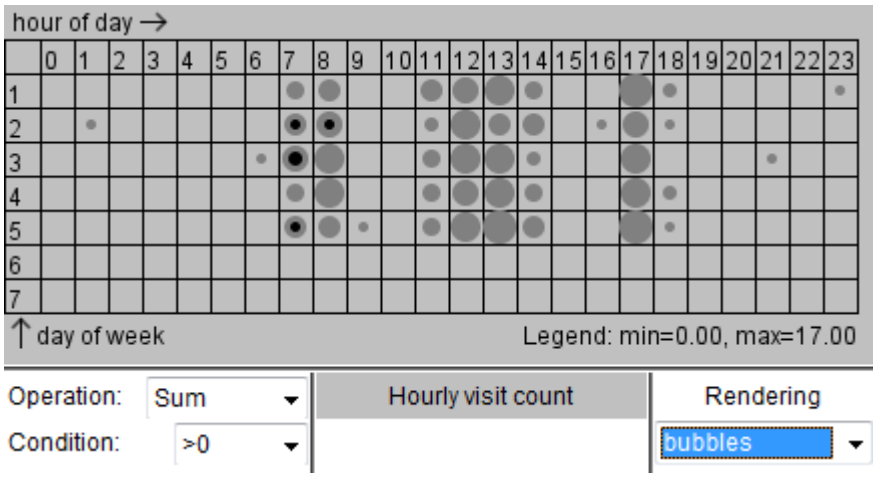

Fig. 12. In the 2d time histogram, the black dots are in the time intervals when the places with the high percentages of the POI type ‘coffee’ were visited (see Fig. 11).

In the bar chart of the POI type occurrences (Fig.11), we see a very high maximal percentage (87.5%) of the type ‘coffee’. Very probably, the set of top ranked places includes one or more coffee bars. We click on the respective bar and observe in the time histogram (Fig. 12) that the stops in this place or these places occurred only in hours 07 and 08, which supports the guess.

It needs to be explained that high proportions of stops labelled by such POI types as ‘coffee’, ‘eating’, or ‘shop’ by themselves do not mean that the places cannot be considered as possible work places. There may be individuals who work in coffee bars, restaurants, or shops. The role of a place for an individual (e.g., whether it is a place to have a cup of coffee or a work place of a barista) can be understood from the temporal pattern of the person’s presence in the place. A work place is expected to have longer time intervals and/or higher frequency of person’s presence than a place visited for the purpose of drinking coffee, eating, or shopping. In our example, we see that the places characterized by the high proportions of the POI type ‘coffee’ are visited only in hours 7 and 8 (Fig. 12, black dots). Hence, it is unlikely that these can be work places of some individuals. Rather, these may be customarily visited coffee bars. Therefore, the place classification with regard to the target meaning ‘work’ needs to be improved, i.e., the scores of the places that are visited only in hours 7 and 8 need to be decreased.

To achieve this, we slightly increase the weight of the cost criterion "% of visits in home time". With the weight 0.4 for this criterion and 0.3 for the two others, we exclude the supposed coffee bar(s) from the set of best scoring places. As a result, we get 34 top ranked places of 34 distinct persons and assign the meaning 'work' to them. For one person, no candidate work place could be found. This may be the same person who visited the area and stayed in a hotel; evidently, he or she had no work place in this area. We refrain from drilling down for investigating the personal data; the knowledge we have got is sufficient for our task.

III.1.3 Interpreting the remaining places

In the further analysis, we consider only those personal places that were visited in at least two different days; otherwise, the information about the place visit times is not sufficient for inferring the place meaning. We filter out 27 places having visits in only one day. Previously, in identifying the home and work places, the attribute "number of different visit days" was involved as a criterion; now, it is used for filtering. Furthermore, we do not use place ranking for identifying places with other meanings than 'home' and 'work'. For 'home' and 'work', we applied ranking based on our background knowledge that almost all people have places with these meanings (roles), and it is typical to have one home and one work place. This reasoning does not apply to places with other meanings. A person may have one, several, or no repeatedly visited shops, restaurants, or bars. Therefore, we use filtering rather than ranking to find places with such meanings.

Based on the list of existing POI types, we expect that the personal places may include regularly visited coffee shops. For finding them, we filter the places according to the proportion of the visits in the morning hours (hours 06-10); see Fig. 13. We find 32 places with proportions about 100%, which belong to 31 distinct persons. We check the selection using the time histogram (Fig. 14) and bar charts of POI types (Fig. 15) and find it quite good; so, we assign the meaning 'coffee' to these 32 places.

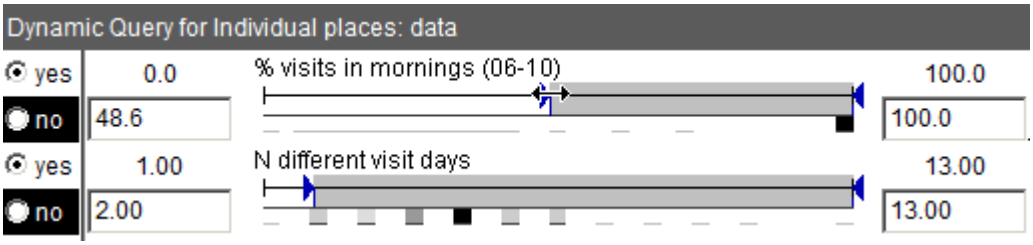

Fig. 13. A fragment of an interactive filtering tool used for the selection of the places visited mostly in the morning hours (06-10).

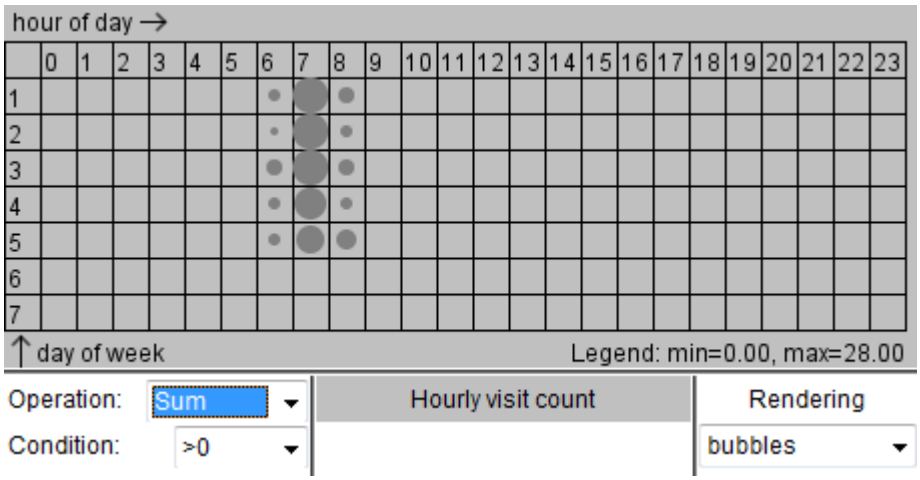

Fig. 14. The 2d time histogram shows an aggregated temporal pattern of stops in supposed coffee places selected by means of the tool shown in Fig. 13.

Fig. 15. The bar charts show the cumulative counts (top) and the maximal proportions (bottom) of the stops labelled by different POI types in the set of supposed coffee places.

To find eating and shopping places, we select places with high values of the attribute "% of visit in lunch and evening times". We assume that eating and shopping places usually include public POIs of corresponding types; hence, these places should have high percentages of occurrences of the POI type 'eating' and 'shop', respectively. Consequently, we use these attributes for filtering and find 63 personal places with the probable meaning 'eating' and 6 places with the probable meaning 'shop'.

After assigning the meanings to these places, we look which POI types still have high maximal percentages of occurrences in the remaining places visited in at least two different days. The only type with a high maximal percentage (30%) is 'hotel'. There are two personal places where the percentages of 'hotel' are about 30%; all others have zero percentages. We select these two places by filtering and see that they belong to two distinct individuals and that they were visited at lunch times of some week days. We refrain from assigning any meaning to these two places, because it is not usual that people may repeatedly visit a hotel in the midday of working days (however, this is a part of the scenario incorporated in the VAST Challenge data).

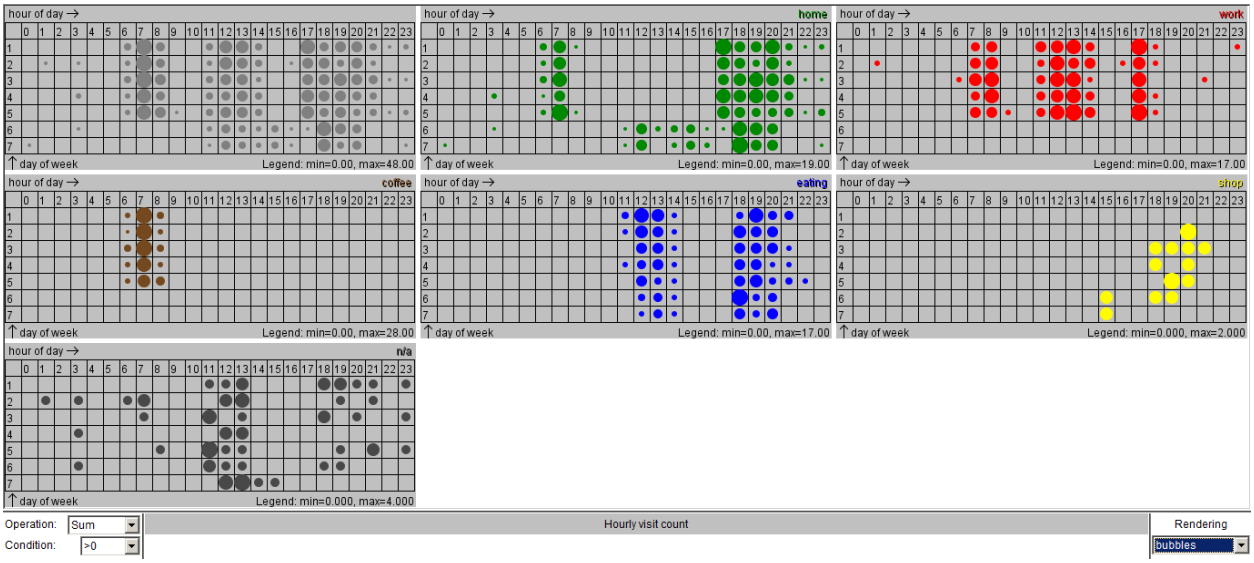

Fig. 16. The 2d time histograms show the temporal patterns of the stop events for different semantic classes of personal places. The histogram in the upper left corner corresponds to the entire set of personal places. The histogram in the lower right corner corresponds to the places the meanings of which could not be identified.

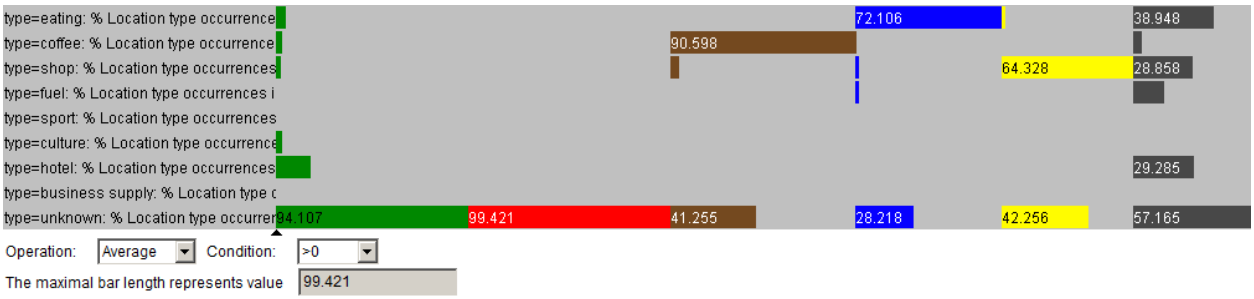

Fig. 17. The multi-attribute bar chart shows the average percentages of stops labelled by each POI type for different semantic classes of personal places.

The final result of our analysis of the personal places extracted from the VAST Challenge data is that we have assigned semantic labels to 170 personal places out of 202, i.e., to 84% of the personal places. The confidence in the meaning assignment is very high, owing to the prominent temporal patterns of place visits (Fig. 16) supported by frequent occurrences of relevant POI types and/or absent or infrequent occurrences of irrelevant POI types (Fig. 17).

The analysis of the 38,225 personal places of 4,286 distinct individuals in the San Diego example was conducted using the same tools and techniques, except that qualitative histograms of land use categories were used instead of the bar charts of POI type occurrences. Since all displays show aggregated data, there is no principal difference between representing tens, hundreds, or thousands of places. Certainly, there are differences between the real San Diego data and artificial VAST Challenge data. A larger number of possible place meanings had to be considered for the San Diego example, including 'transport', 'education', 'religious facility', 'fitness', and others. We assumed that some people might have two homes or two work places and classified some places as second home or second work. The temporal patterns of place visits were not so "clean" and easily interpretable as in the VAST Challenge example; therefore, the confidence in the meaning assignment was lower than in the VAST Challenge case.

We managed to attach meanings to 65% of the San Diego places. 3,873 persons (90.4% of all) have got home places, and 695 of them have got places with the meaning 'second home'. We could identify probable work or study places only for 2,171 persons (50.7% of all); for 529 persons, we found probable second work places. For 1,950 persons (45.5%), it was possible to find both home and work places. The largest class of personal places is 'shopping' (4,695 places), other large classes are 'eating' (2,194), 'social life' (1,497), which includes places with many visits in the evening and night hours and on the weekend, and 'transport' (1,315).

Please note that, although we analysed personal places, the whole analysis in both case was done without seeing any personal data. We used only aggregated data and information about the number of currently selected places and the number of persons they belonged to. Hence, our experiment has shown that it is possible to determine meanings of personal places without seeing personal data and violating personal privacy.

III.2 Analysis of public places

In the VAST Challenge example, we have 41 public places extracted earlier from the episodic trajectories. A place was selected as public if it was visited by at least 2 distinct persons (the threshold was low because there are only 35 persons in total). From the description of the challenge, we know that all people work in the same company. Hence, we can expect that one of the public places corresponds to this company. We identify it using the ranking tool for public places with criteria "total number of visit-days", "% of visits in work time", and "% of visits in home time"; the first two are maximized and the third one is minimized.

For other possible place meanings, we cannot assume that there may be only a single place with each meaning. Therefore, we analyse the places using filtering rather than ranking, as we did previously for the personal places. We identify coffee shops, eating places, and shops in the same way as with the personal places. We detect a place with 100% of POI occurrences of the type 'business supply' and assign the meaning 'business supply' to it. Analogously, the places with high percentages of occurrences of the types 'fuel', 'sport', 'culture', and 'hotel' receive these meanings after checking their compatibility with the temporal patterns of place visits. In this way, we have labelled 24 places. For the remaining 17 places, almost all stops have unknown POI types; hence, we cannot rely on the POI information anymore. We can guess about the place meaning only on the basis of the temporal distributions of the stops.

We select places that were visited only on weekend. Among the unlabelled places, there is only one such place. This cannot be a church, because the visits on Saturday span from 10 to 16, and

there is also a visit in hour 18. This may be a place for some kind of recreation, such as a park, where people are not expected to pay money (no credit card transaction records could be associated with it). We assign the meaning ‘recreation’ to this place.

We guess than the remaining 16 places may include home places of some people. These may include multi-family buildings where several people live, or common parking places, where people leave their cars while they are at home. Besides, if some persons were visited by others, their home places might be included in the set of public places. Therefore, we look if there are places with high percentages of visits in the home time intervals, i.e., from hour 18 till hour 08 on the working days and the whole weekend. We find 11 places with more than 70% of visits in these times. The summarized temporal pattern of place visits in the 2D time histogram looks like a home pattern; however, the selected subset may include places that were just occasionally visited in home times. We look at the values of the attribute “N visit-days total” and see that the smallest number among the selected places is only 2. The next smallest value is 11, which is sufficiently high, taking into account that the data cover a period of only 14 days. We exclude the place with 2 visit-days and assign the meaning ‘colleague’s home’ to the remaining 10 places.

6 public places still remain unlabelled. In the 2d time histogram for these places, we see that there were many stops in hour 11. To select the places visited in this hour, we compute an attribute “% visits in hour 11”. The values of this attribute range from 0 to 100, the second smallest value after 0 is 33.3%. There are 5 places with such high proportions of stops in hour 11. Their joint temporal pattern of stops looks very regular, which should have a certain meaning. Since we cannot guess what the meaning is, we make a special category ‘hour 11 place’ including these particular places. Finding these particular places corresponds to the VAST Challenge scenario.

Finally, only one public place remains unlabelled. It was visited only twice, which does not give us enough information for determining its meaning.

The final result of assigning semantic categories to the public places is presented in Fig. 18 (the temporal patterns of the stops) and Fig. 19 (the average percentages of stops labelled by the existing POI types).

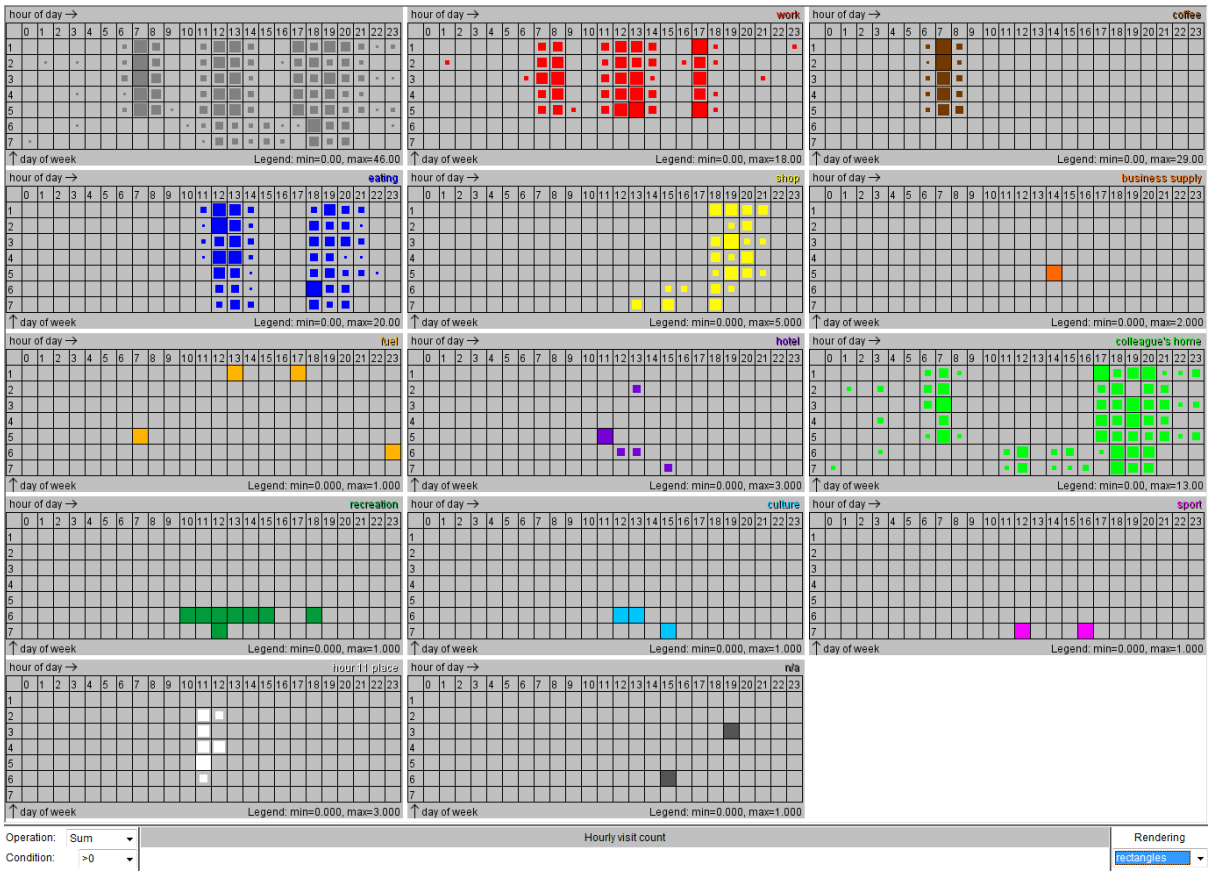

Fig. 18. Temporal patterns of stop events for different semantic categories of public places.

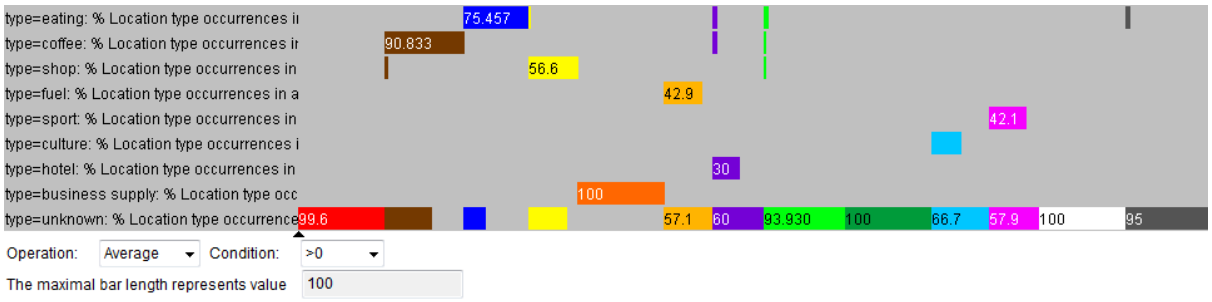

Fig. 19. Percentages of stop events labelled by the available POI types for different semantic categories of public places.

In a similar way, we analysed 9,301 public places in the San Diego case, involving land use data instead of the counts and percentages of POI type occurrences. This required more effort, since the land use classes are much more numerous than the POI types in the VAST Challenge example. Another complication was that the temporal patterns of the visits to the real public places were much more blurred than those for the artificial places. The reason may be that many real public places may have multiple uses; for example, shopping centres may include restaurants, bars, cinemas, and fitness rooms. We were able to assign semantic labels to 5,144 public places (55.3%).