Dealing with Multiple Source Spatio-temporal Data in Urban Dynamics Analysis

João Peixoto[1] and Adriano Moreira[2]

Mobile and Ubiquitous Systems Group, Centro Algoritmi,
Universidade do Minho, Guimarães, Portugal
peixoto@kanguru.pt, adriano.moreira@algoritmi.uminho.pt

Abstract. Capturing, representing, modelling and visualizing the dynamics of urban mobility have been attracting the interest of the research community recently. One of the drivers for recent work in this area is the availability of large datasets representing many aspects of the urban dynamics. Applications for these studies are diverse and include urban planning, security, intelligent transportation systems and many others. Quite often, the proposed approaches are highly dependent on the data type. This paper describes the definition of a set of basic concepts for the representation and processing of spatio-temporal data, sufficiently flexible to deal with various types of mobility data and to support multiple forms of processing and visualization of the urban mobility. A place learning algorithm is also described to illustrate the flexibility of the proposed framework. Available results obtained by the integration of geometric and symbolic data reveal the adequacy of the proposed concepts, and uncover new possibilities for the fusion of heterogeneous datasets.

Keywords: Urban modelling, space-time dynamics, data fusion.

1 Introduction

The mobility of citizens in an urban area is the source of various problems: traffic congestion, environmental impacts, inadequacy of public transport, and spreading of diseases, among others. For this reason it is important to understand the mobility behaviour of individuals in space, understand space itself, and understand the use people make of the urban space as a way to reduce and possibly eliminate these difficulties.

The dynamics associated with the mobility in urban areas always has two components, Time and Space, rising new challenges on how to capture, represent and visualize these dynamics. While capturing the presence and mobility of people in urban spaces has evolved enormously in recent years, movement representation and visualization still faces many challenges. Actually, the huge size of datasets being collected these days is creating more challenges to representation and visualization rather than solutions (in spite of their great potential for mobility analysis).

As referred by Yu and Shaw [1] the current Geographic Information Systems (GIS) are structured to represent the spatial component of data but lack good support for the temporal component. For this reason, some authors have developed platforms that

B. Murgante et al. (Eds.): ICCSA 2012, Part II, LNCS 7334, pp. 450–465, 2012.

provide support to spatio-temporal data, such as SECONDO [2], with the aim of representing the time dependence of mobile artefacts through the use of abstract definitions to represent the positions or shapes of objects over time and space, respectively [3]. On the other hand there are also challenges in how this information could or should be presented for human movement analysis. Thus, several studies presented recently explore different forms of visual representation of movement data, such as the iSPOTS project [4] which illustrates the occupation of space or, for example, the visualization of human travel behaviour based on the trajectory of money bills [5].

However, regardless of how the dynamics of an urban space is represented, most of these works focus only on one type of mobility data. Although our research work is focused on the analysis of the dynamics of urban space, our initial aim is to create a flexible and comprehensive conceptual framework for the representation of movement processes that allows the same concepts to be applied to different types of data from different sensors, such as GPS, Wi-Fi, GSM, ticketing systems, as well to the different modalities of urban mobility. Our approach to capture the dynamics of the urban space is based on merging the individual mobility profiles of people. This approach aims to benefit from the current capability of smartphones and other personal devices to be used as proxies to observe human spatio-temporal behaviour. The first step in the analysis of the urban dynamics is, then, the automatic creation of personal mobility profiles from multi-sensor data.

The next section in this paper describes some of the work developed in the field of analysis and visualization of urban mobility. Section 3 describes the concepts that are the basis of our proposed framework for the representation of spatial-temporal data. In Section 4, data from three types of sensors is mapped into the proposed concepts and three of the major transformation processes are described. Finally, in Section 5, some conclusions and open questions are discussed.

2 Related Work

Recently, several studies have been presented in the area of visualization of urban mobility dynamics, using different techniques. One of these techniques analyse urban mobility using the temporal variation of the occupation that individuals make of the urban space [4] [6]. This type of representation is based on the creation of temporal snapshots of space occupation. However, due the dynamics of the urban space, this approach may not be the most appropriated for the analysis of pattern changes [7]. Another problem is the definition of mobility in these approaches, because they represent the mobility through the variation of space occupation over the time and not the real movement of individuals. The analysis of mobility based on the use of space does not allow the extraction of more depth conclusions about the urban mobility. Thus, the application of this approach may be useful for the planning of urban space based on the detection of concentration areas of individuals, but not adequate for the detection of problems caused by mobility itself, such as traffic congestion.

Deep understanding of the phenomena associated with urban mobility involves the visualization of trajectories and flows, which truly reflect the movements of individu-

als. In this area one of the approaches to represent trajectories is with vectors [8]. Through this approach it is possible to have some sense of mobility, since it allows the representation of an individual according to a spatio-temporal reference. However, as this representation is based on the observation of the instantaneous movement, the simultaneous perception of the origin and destination of the movement is not easily transmitted. Alternatively, some researchers are using a different technique for representing paths through interconnected source-destination pairs [5]. Although with this approach it is possible to visualize the trajectories, several questions arise regarding the outcome of the visualization. First, if the interval between samples is large, intermediate movements are lost and then trajectories become twisted, as such some behaviours are not represented. Second, to connect the source to destination we may have to affect the Time component, since the analysis is not done continuously, but by time intervals, consequently losing this representation the notion of space change over the time and creating the same issues associated with representation by snapshots. In our research work we intend to study these and other issues related to the visualization of mobility through the representation of trajectories so as to explore new paradigms for the representation of mobility. Our approach is based on abandoning the snapshot representation of artefacts (individual or object), and create personal and global maps of mobility. To achieve this goal it is important to, first, properly structure the information in order to have a conceptual framework for the representation of mobility in an urban area and verify what type of data match with this structure.

3 Concepts for Movement Representation

The work that we have done so far defines eight general concepts that characterize our conceptual framework for the representation of mobility of an individual artefact (Figure 1). These concepts are designed to fit the data since its acquisition stage, until we get homogeneous representations of the major movement processes, be they of a single individual or of a group of individuals.

3.1 Raw Data

It all starts with the data collected by a multitude of sensors about the movement of an artefact. Until recently, GPS receivers have played a major role in data collection about movement. Fortunately, recent advances in mobile devices created the possibility to collect large amounts of data about the mobility of their users. These devices not only support the collection of geo-referenced data through their integrated GPS receivers, they also enable the collection of data about the use of Wi-Fi and GSM networks, the detection of nearby devices (persons), the logging of data generated from accelerators, and much more. Urban infrastructures are also contributing to these increased sensorial capabilities. Among others, public transportation operators often make use of ticketing systems that collect data about people entering or leaving buses and metro stations. Public authorities also collect data about the intensity of traffic flows across a particular street segment.

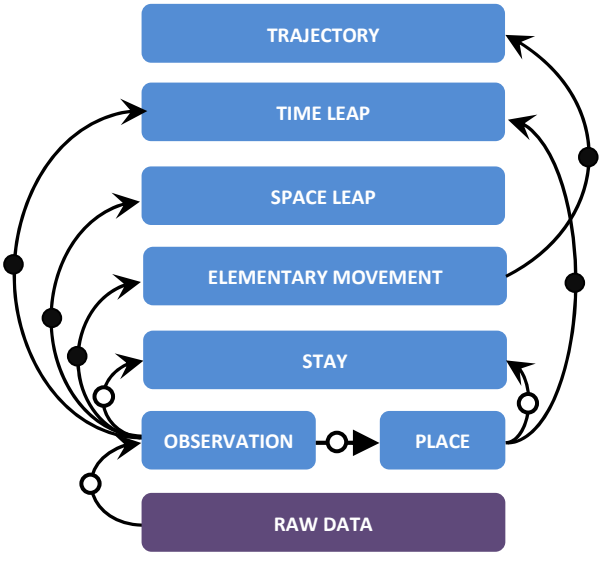

Fig. 1. Layered Structure of the Information Concepts

The result of this technological equipped urban environment is, from the mobility analysis point of view, a giant sensor network producing data in huge quantities. These datasets are, though, very heterogeneous: on the formats used to represent the mobility data; on the position precision and accuracy; on the time stamping accuracy and synchronization; on the sampling rate; on the spatial reference system being used to represent points or places in space.

The heterogeneity of the spatial reference model is of major importance when fusing data coming from different sensors. The two major spatial referentials are those based on the WGS-84 datum, and those based on addresses or names of places. The first one is a geographic space model, while the second is a symbolic one. It is on the second type where we find the greatest diversity: postal addresses, postal codes, networks cells identifications, bus stops identifications, etc.

Throughout the remaining of this section we propose a set of concepts for the representation of the mobility processes, in an attempt to merge most of the above mentioned heterogeneity into a more generic and homogeneous reference model.

3.2 Observation

A mobile artefact can be observed from different perspectives through different sensors. Each of these sensors can collect information on the artefacts with different attributes or features. As these raw datasets include different attributes, some form of normalization is required. The *Observation* concept aims to realize the first step in this normalization process, defining a basic set of information necessary to characterize the observation of an artefact from the mobility point of view.

Regardless the sensor used to acquire the mobility data, these data reflect the observation of a phenomenon. The *Observation*, and can be described abstractly as:

The observation of an artefact in a specific point of a spatio-temporal space

and is represented by the following attributes:

(Id_Observation, Artefact, Location, Timestamp)

The first attribute allows distinguishing the observation of the same artefact in the same place and time instant by multiple sensors. The Artefact attribute uniquely identifies the observed artefact. The Timestamp attribute records the time instant of the observation. The Location attribute describes, with variable levels of detail, the location where the artefact was observed. Since the raw data obtained from different sensors can define the location of an artefact according to a geometric or symbolic referential, the Location concept combines both geometric and symbolic representations. Geometric, when the representation of Location is based on cartographic coordinates. Symbolic, when the Location is characterized by the description of its location, for example a building name, an address, or the ID of a cell in a GSM network.

3.3 Location vs. Place

Regardless the type of sensor used to characterize an *Observation*, each *Observation* describes an artefact as a point in spatio-temporal referential (as shown in Figure 2a), i.e., the artefact is described by the instant time and space dimension.

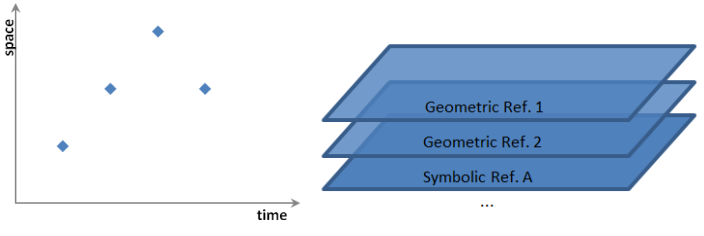

Fig. 2. Observations in a spatio-temporal referential (a), and the multitude of the spatial dimension (b)

For dimension time there is an almost universal definition and referential and, as such, the *Timestamp* attribute definition of an *Observation* does not cause major difficulties. The same does not apply with the spatial dimension, where such uniformity does not exist. As discussed earlier, space may be described in a geographic or symbolic manner and using several independent referentials (Figure 2b).

We define *Location* as a point in an arbitrary spatial referential, geometric or symbolic, or both, and represent it as: ***location = {symbolic name, position},*** where position is represented by a pair of coordinates such as:

position = (latitude,longitude) or: ***position = (X,Y)***

In turn, the aggregation of one or more nearby *Locations* defines a *Place*. The *Place* is not so much a specific point in one space dimension, but one region of space defined by the aggregation of *Locations*, whether they are symbolic or geometric.

The necessity to aggregate *Locations* emerges from the following reality. Considering that the location of an artefact in a sequence of *Observations* is described geometrically (for example, by pairs of GPS coordinates), an immobile artefact might be reported to be observed in different points of the space domain, in consecutive *Observations*, due to the sensor noise. Similar phenomena are observed in the symbolic domain. However, the symbolic space is usually a discrete one, while the geometric space is continuous, making this last one more prone to sensor noise. This way, even if there are slight changes of position when we conduct the process of aggregation and the subsequent detection of a *Place*, we get a better geographical characterization of the artefact and consequently a smaller number of distinct places visited by the artefacts. Extracting *Places* from sets of *Observations* might require different processes for geometric and symbolic data. Although many approaches have been described in the literature for both domains, we assume that combined approaches are still an open research challenge.

3.4 Suspension of Movement

Based on consecutive observations $\{o_{i-1}, o_i\}$ of an artefact over time, the artefact can be in one of two possible states: a moving state or a stationary state. The artefact is in stationary state when in the most recent observations the *Place* (P) is the same, no matter the *Location* (l).

$$\left(l_i \in P_j\right) \wedge \left(l_{i-1} \in P_j\right) \tag{1}$$

Stays vs. Time Leaps

Associated with the stationary state of an artefact, we define the concept of *Stay* as:

Time interval between the first and last observation of an artefact in the same place

and represent it by the following attributes:

(Id_Stay, Artefact, Place, Timestamp_Initial, Timestamp_Final)

Thus, we assume a *Stay* when for several consecutive observations of the artefact, the *Place* is the same. In the case of symbolic representations of *Place*, the transformation process that extracts *Stays* from *Observations* is simpler, as opposed to geometric representations, due the fluctuations of the positioning information. Several authors have already addressed this problem of detecting stays (also referred as stops, or regions of suspension of movement), in particular for the case of GPS traces.

It should be noted, however, that there are two formal restrictions to the definition of *Stay*. First, the time interval must be longer than zero. Second, the largest time interval between two consecutive *Observations* (from where the *Stay* emerged) should be short enough so that we can assume that the artefact did not leave the correspond-

ing *Place* between *Observations*. Otherwise, if there is a long time gap between *Observations*, we cannot derive a *Stay* since it is not reasonable to infer intermediate observations (along a straight line connecting the two observations in the spatio-temporal space). These situations occur when, for example, a GPS tracked artefact enters a building. In order to address these situations we defined a concept of *Time Leap*, described as:

Long time period between two sequential observations of an artefact in the same place

and represented by the following attributes:

(Id_TimeLeap, Artefact, Place, Timestamp_Initial, Timestamp_Final)

Note that, the representation of a *Time Leap* and that of a *Stay* are identical, while representing different concepts with different semantic meanings.

3.5 Movement

The artefact is in the moving state when there is a change of *Location* (l) and the two most recent observations are not in the same *Place* (P), (one or both observations might even not be associated to a *Place*).

$$(l_i \neq l_{i-1}) \wedge \left[\forall_j \left(l_i, l_{i-1}\right) \notin P_j\right] \tag{2}$$

Elementary Movement vs. Space Leap
In the movement state, we expect to observe variations in the *Location* attribute over the time, so there must be a concept that represents this variation. We name it *Elementary Movement*, and describe it as:

A Change of Location of an artefact occurred over time

represented by the following attributes:

(Id_Movement, Artefact, Location_Start, Location_End, Timestap_Initial, Timestap_Final)

We only consider that an *Elementary Movement* exists when, for a pair of observations at different locations in consecutive time instants, it is reasonable to interpolate the intermediate positions at intermediate time instants, i.e., only in situations where we can assume that the artefact moved from the initial *Location* to the final *Location* along a straight line at constant speed. This is often assumed when transforming a set of consecutive GPS records into a line representing the artefact trajectory.

For example, if a pedestrian is observed based on its GPS trace and these observations have been acquired at time intervals of one second, it is reasonable to infer that the artefact has realized elementary movements, as it is also reasonable to infer the intermediate position between two observations. On the other hand, if the time interval between observations is in the order of one minute, one cannot easily derive *Elementary Movements* from pairs of consecutive observations, because the space that can be travelled in such time interval is significant and we cannot assume that the artefact has actually travelled along a straight line. On the other hand, sampling the geographic position of a flying airliner at one minute intervals might be enough to derive *Elementary Movements*.

In situations where the time interval between consecutive observations of the artefact does not allow inferring with precision the intermediate positions of the artefact due to the large space of possibilities, we cannot consider that an *Elementary Movement* occurred, but instead a *Space Leap*. The *Space Leap* is described as:

A Change of Location of an artefact occurred over a long time period

represented by the following attributes:

(Id_SpaceLeap, Artefact, Location _Start, Location _End, Timestap_Initial, Timestap_Final)

Note that the representations of a *Space Leap* and of an *Elementary Movement* are similar, but their semantic meanings are different.

Trajectory

Finally, at the top layer of our conceptual framework we have the concept of *Trajectory* that represents a set of *Elementary Movements* ordered in time for the same artefact. The concept of *Trajectory* is described as:

Time-ordered list of Elementary Movements of an artefact over the space

and represented by the following attributes:

(Id_Trajectory, Artefact, List of Elementary Movements)

In turn, the set of existing trajectories at a given spatio-temporal interval for a given artefact or group of artefacts might lead to the representation of flows that exist in the urban space, the final goal that we aim to achieve with our study.

4 Mapping Real Data into the Proposed Framework

In order to validate the concepts of our proposed framework for the representation of spatio-temporal data, it is important to realize how well real data obtained through different sensors (different raw data) match the concepts in the framework. This task also triggers the design of a set of transformation processes responsible for the selection and adaptation of existing information, and also, to infer new information based on existing data.

In this section we take a data set comprising a set of RAW records collected by one single user, and describe how these data is mapped into our conceptual framework. Our focus is on the concepts of Observation, Place and Stay, and on the processes used to derive Places from Observations, and Stays from Observations and Places.

This data set includes records obtained from three types of sensors: GPS, Wi-Fi and GSM. By describing the transformation processes, we propose an approach to data fusion, where the three types of records are processed simultaneously to extract Places and Stays.

4.1 The Data Set

Many of the previous works presented in the area of visualization of urban mobility rely on collections of readings obtained from GPS receivers. These records include

data beyond the position, such as the information of the temporal moment in which the position was acquired (a timestamp), the speed, the orientation (bearing), and other attributes. Since this type of mobility information has frequently been used in the study of urban mobility and it is easy to obtain, our framework should also adapt to this type of mobility data. On the other hand, resorting to GPS data to represent human mobility raises some concerns. The first one is about temporal coverage since it is difficult to collect GPS data indoors, and humans spend more than 80% of their time indoors. Therefore, most of the observation period is not actually observed. The second problem is that collecting GPS data in real world situations is technically difficult. One common approach is to give people a GPS receiver and a data logger to carry on during a specially prepared experiment (eventually disturbing the real daily live routines). The limitations of this approach are the limited number of persons observed, and the short observation period. An alternative is to rely on people's smartphones. Most of the recent smartphones integrate multiple sensors, and these sensors can be used to observe people spatio-temporal behaviours in many ways.

In our experiments, we resorted to the smartphone approach, aiming to overcome the limitations of the GPS only solution. By asking a group of people to install a small application on their Android smartphones we have been able to collect data during long periods of time and from multiple sensors: GPS coordinates, the nearby Wi-Fi Access Points, and the nearby GSM cells. After a few days, people using this application forget its use, and do not constrain their movement behaviour – they do not feel being involved in an experiment. The second advantage is that we collect data both outdoors (the three types) and indoors (Wi-Fi and GSM).

The data used throughout this section, for illustration purposes, were collected for a single user and over several months. Many other users were involved, but their data is not used here. The data reflect the mobility of a person in his daily normal activities. The following tables illustrate the raw data that were collected.

Table 1. Raw data collected from the GPS sensor

Timestamp	Latitude	Longitude	Altitude	Speed	Accuracy	Bearing
2011/06/29 15:25:07	1,297077	103,7808	93,5	0,75	17,88854	65
2011/06/29 15:25:18	1,297077	103,7808	108,2	0,75	26,83282	162,4
2011/06/29 15:25:31	1,297213	103,7806	134,4	1	40	283,8

Table 2. Raw data collected from the Wi-Fi sensor

Timestamp	BSSID	RSSI	SSID
2011/06/29 15:25:08	00:27:0d:07:d6:c0	-90	NUS
2011/06/29 15:25:11	00:27:0d:07:d6:c0	-88	NUS
2011/06/29 15:25:12	00:27:0d:07:d6:c0	-88	NUS

Table 3. Raw data collected from the GSM sensor

Timestamp	CID	LAC	MNC	SIGNAL_STRENGTH
2011/06/29 15:25:08	962335	441	3	9
2011/06/29 15:25:10	962335	441	3	8
2011/06/29 15:25:11	962335	441	3	8

The Wi-Fi raw data include a timestamp, the BSSID (MAC address that identifies the AP to which the artefact is currently connected), the RSSI (received signal strength indicator), and the SSID to identify the network to which the artefact is connected (other attributes were collected but are not represented here). The GSM raw data includes a timestamp, the cellID, the Location Area Code, the Mobile Network Code, the received signal strength, and other attributes not represented here.

4.2 Mapping RAW Data into Observations

Mapping GPS, Wi-Fi and GSM raw data into a set of *Observations* is straightforward. However, the transformation processes must be specific for each type of raw data and, in this example, different for GPS and Wi-Fi/GSM. For GPS data, we mapped the raw timestamp into the *Observation* timestamp, and the pair of coordinates into the *Observation* Position. For the Wi-Fi data, we mapped the raw timestamp into the *Observation* timestamp, and the BSSID into the Location. For GSM, we mapped the raw timestamp into the *Observation* timestamp, and the cellID into the Location. No data cleaning was performed. The table 4 illustrates the resulting set of *Observations* (the identification of the artefact was not included, and the timestamp was simplified).

4.3 Extracting Places from Observations

Automatically detecting Places that are relevant for one single person (e.g. the workplace) or for a group of persons (e.g. a popular place at a certain urban location) is an activity known as "place learning". Many approaches for place learning are described in the literature, most of them dealing with a single type of observations at a time, like GPS or GSM. Recent work in this field is addressing place learning by integrating observations from multiple sensors, such as GPS, Wi-Fi, and accelerometers, collected using smartphones [9]. In this work we propose a method for place learning based on a probabilistic model applied to observations obtained from GPS, Wi-Fi and GSM sensors. The novelty of this method comes from the simultaneous processing of the three types of observations, thus performing data fusion while clustering the observations to identify places. The proposed approach is an alternative to density-based spatial clustering algorithms.

Table 4. A set of *Observations* obtained from GPS, GSM and Wi-Fi raw data sets (the *Observations* are sorted chronologically)

Timestamp	Location			Optional Attibutes
	Position		Symbolic Name	Sensor_type
	Latitude	Longitude		
15:25:07	1,297077	103,7808		GPS
15:25:08			00:27:0d:07:d6:c0	WIFI
15:25:08			962335	GSM
15:25:10			962335	GSM
15:25:11			00:27:0d:07:d6:c0	WIFI
15:25:11			962335	GSM
15:25:18	1,297077	103,7808		GPS

Density-based spatial clustering algorithms, such as DBSCAN [10] or SNN [11] define the similarity between multidimensional points using a distance function. For datasets where each point is described by a pair of coordinates, the most popular distance function is the Euclidean distance (2-norm). In our case, while GPS observations include a position (pair of coordinates) in their description, Wi-Fi and GSM observations do not. Therefore, a different distance function must be defined. Our approach to define such a distance function relies on the basic concept of *Observation*: one Observation describes what a particular sensor measured at a particular time instant and location, and not the location itself. Consider the example of a GSM based observation: the observation states that, at a particular time instant, the strongest GSM cell in the neighbourhood was C, and not that the observation was taken at the location of the C cell tower. Since GSM cells are often quite large, the consequence is that two samples of the GSM sensor taken at two far apart locations might be similar. Therefore, when trying to learn places from sets of observations, these two samples (observations) might end up being part of different places, meaning that cell C was "visible" from these two places.

We model the above described concept through a distance function that describes not the Euclidean distance between points in the dataset but the probability that two points (observations o_i and o_j) having been taken at the same place:

$$P_{sameplace}(o_i, o_j) \tag{3}$$

Dealing with three different types of observations simultaneously requires the use of 6 different probability functions:

Table 5. Probabilities that two observations have been taken at the same place

Prob. function	GPS	Wi-Fi	GSM
GPS	P_1	P_2	P_3
Wi-Fi	P_2	P_4	P_5
GSM	P_3	P_5	P_6

For the probability P_1, between two GPS observations, the Euclidean distance is a good indicator. If two observations are geometrically close, then they probably refer to the same place, and, therefore, P_1 can be described as:

$$P_1(o_i, o_j) = e^{-ED(o_i,o_j)/R_1} \tag{4}$$

where *ED*() is the Euclidean distance between observations o_i and o_j, and R_1 is a parameter that relates the Euclidean distance to the closeness of the two observations. Note that P_1 takes a value of 1 for two observations taken at exactly the same position, and tends to 0 as the Euclidean distance goes to infinity.

For P_2 and P_3, the geometric distance cannot be used since Wi-Fi and GSM observations are described by symbolic locations. The same applies for P5 (Wi-Fi - GSM), since both observations are described by symbolically. In these cases we rely on the time difference between the observations: two samples taken within the same short time interval must refer to the same place. Therefore, P_2, P_3 and P_5 are defined as:

$$P_k(o_i, o_j) = e^{-|t_i - t_j|/R_k}, k = 2, 3, 5 \quad (5)$$

where R_2, R_3 and R_5 are parameters that relates the time difference between the observations and the closeness of the two observations.

While estimating the closeness of two Wi-Fi observations, the characteristics of the Wi-Fi networks must be taken into account. The coverage area of one single Access Point (AP) is typically small and assumed to be a circle with a radius of 50 meters or less. Therefore, if two observations refer to the same AP, one can assume that they were taken from the same place. If they refer to different APs, then the time proximity can be used as in (5). So, for pair of Wi-Fi observations, P_4 is defined as:

$$P_4(o_i, o_j) = \begin{cases} P_{sameAP}, & AP_i = AP_j \\ e^{-|t_i - t_j|/R_4}, & AP_i \neq AP_j \end{cases} \quad (6)$$

where P_{sameAP} is the probability of two samples referring to the same place given that the observed AP is the same. We do not set this probability to one because in places with poor coverage of Wi-Fi networks, the same AP might be detected from different nearby places. By setting this probability to a lower value (we are using 0.975), we do not limit the place size to the typical Wi-Fi cell size.

For the closeness of two GSM observations, also symbolic, the model used for Wi-Fi cannot be used since GSM cells are typically much larger in coverage area. Here we resort to the temporal proximity, but weighting differently depending if the two cells are the same or different:

$$P_6(o_i, o_j) = \begin{cases} P_{sameCell} \times e^{-|t_i - t_j|/R_6}, & cell_i = cell_j \\ P_{difCell} \times e^{-|t_i - t_j|/R_6}, & cell_i \neq cell_j \end{cases} \quad (7)$$

the parameters $P_{sameCell}$ and $P_{difCell}$ being used to weight the probability.

For building the places, an iterative algorithm is used, where each new observations is added to one of the existing places if the probability of being taken at one place is higher that a predefined threshold (P_{min}). Otherwise, the observation is used to create a new candidate place. If P_{min} is exceeded, the observation is added to the place with higher probability.

One place is described by its GPS part, the Wi-Fi part, and the GSM part. The GPS part is represented by the centroid of all the GPS observations that have been added to the place, and the timestamp of the most recent GPS observation added to the place. The Wi-Fi part is described by the BSSIDs of all Wi-Fi observations that have been added to the place, and the timestamp of the most recent Wi-Fi observation added to the place for each BSSID. A similar representation is used for the GSM part.

The probability that an observation has been taken at a given place is the highest of the three probabilities that compare that observation with the three parts describing a place. Since each new observation that do not exceed P_{min} is used to start a new candidate place, and since most of the observations collected while the person is moving do not exceed P_{min}, a large number of candidate places are created by the algorithm. A candidate place is assumed to be a real relevant one if the total accumulated time spent at that place is longer than a minimum amount of time (e.g. two minutes).

Figure 3 illustrates the results of using the above described algorithm to detect the places visited by one single person during a one month period. A total of 13 places, with more than 2 minutes of total staying time have been detected. The place in red is the most relevant one, with a total staying time of 402,4 hours (in one month). Note that the algorithm has been able to distinguish between different places in very close locations (inset in Figure 3).

Fig. 3. Places detected from one month of data (159380 observations, from which 92991 are GPS, 60427 are Wi-Fi, and 5962 are GSM)

The detailed assessment of the quality of the place learning approach here described is out of the scope of this paper due to space limitations. However, a validation has been performed by comparing the detected places with a diary of the person under observation. All the 13 detected places are actually places that have been visited. Therefore, the false positives rate is 0%. On the other hand, not all the relevant places registered on the diary have been detected. In a few cases, this was due to the lack of observations (failure in the acquisition process). Other cases were due to the fact that only GSM observations were collected during the stay at those places, and those observations were too sparse in time to be grouped into one place.

4.4 Extracting Stays from Places and Observations

A place, as described in the previous section, is completely described by the set of observations that were clustered to create it. Therefore, computing the stays at each place from the place description is straightforward. In the following analysis, we assume that a stay occurred whenever the time elapsed between consecutive observations in a place do not exceed a given threshold (T_{max}). By concatenating all the consecutive time intervals that do not exceed T_{max}, one detects the stays at a given place. Figure 4 represents the stays (black lines) extracted from the set of 13 places shown in Figure 3. In Figure 4, the blue dots represent the GPS observations (y represents the

distance from a reference point), the red dots represent the Wi-Fi observations (y represents different BSSIDs), and the green dots represent the GSM observations (y represents different cells). The inset in Figure 4 shows the details for one single day. This example shows that most of the time (stays) is assigned to one of the 13 places (56,5% of the total time in one month). Figure 4 also shows that there are temporal gaps between stays. These gaps represent the periods of movement, the periods where there is no data (inset in Figure 4), and the periods where the observations are too sparse to be grouped into a place.

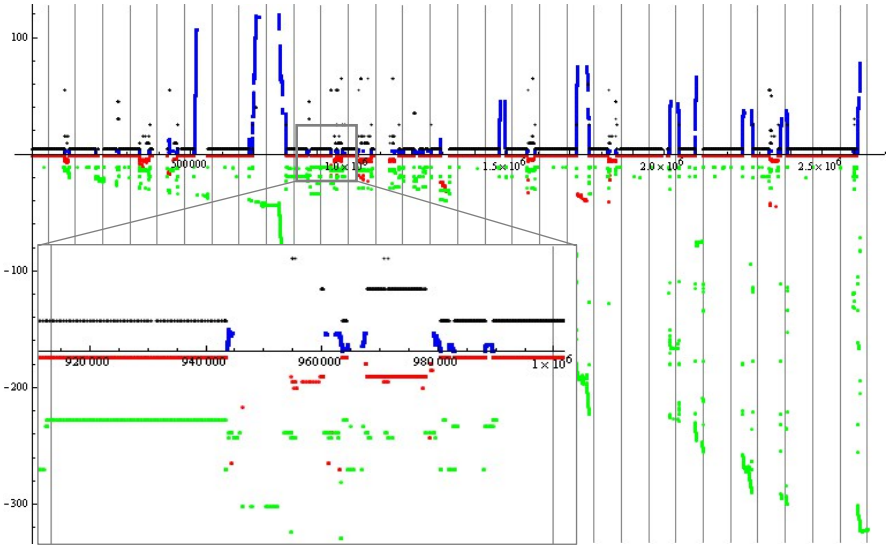

Fig. 4. Stays extracted from the observations in each place (T_{max}=60 seconds)

Stays represent a fundamental concept for the characterization of places. The results in Figure 4 uncover relevant information about the time profile of each place and, consequently, about the importance and relevance of that place for its tenant. Similar information about public places can be used to understand how particular urban areas are used by persons, and how the use relates to local events or availability of infrastructures. In our particular case, stays are of major importance for the characterization of transportation requirements in urban spaces, as the time profiles of places can be used to distinguish residential from commercial or industrial areas.

Stays are also the basis for the detection of origin-destination trajectories. A description of that process, as well as the description of the other mapping processes identified in Figure 1 is left for future articles.

5 Discussion and Conclusions

The mapping of multi-sensor data performed in this study allowed us to verify that the proposed concepts are appropriate to represent the three types of records used. It also

supported the identification and design of various transformation processes required to map the data into the proposed concepts.

As illustrated in section 4.2, the proposed structure for an *Observation* can be used to represent both symbolic and geometric positions. The unified representation for the observations enabled the design of a generic process to infer *Places* while performing fusion of multi-sensor data. Extension of the proposed place learning algorithm to accommodate other sources of data only requires the definition of the probability functions that measure the probability of pairs of observations being taken at the same place. The results in section 4.4 reveal another benefit of using a unified representation for observations by illustrating how simple is the process that extracts *Stays* from the set of *Observations* that describe each *Place*.

Mapping *Places* and *Observations* into the remaining concepts identified in Figure 1 demands the design of other transformation processes. Besides these transformation processes, additional concepts might also need to be defined. Among them is a generalization of the *Trajectory* concept as our notion of *Trajectory* is only linked to the concept of *Elementary Movement*, i.e., we only consider a *Trajectory* exists when it is possible to infer intermediate positions between observations. As such, we are not covering the situations where *Space Leaps* occur while going from one place to another. One possible solution to this problem might be to define a new type of trajectory as a sequence of *Space Leaps* between places (eventually merged into an origin-destination trajectory). Currently our work is focused on the validation of the proposed concepts and transformation processes using a variety of datasets, including data from transportation systems (e.g. ticketing data used in buses).

Other challenges in this context are related to processing massive datasets. One of the difficulties already identified with the mapping we have been conducting is related to the space occupied at the level of storage in the database system of the observations data. Because we are working with individual data without any kind of aggregation, transformation processes need to deal with a large number of records (for example, the dataset used in section 4, representing one single user for a period of one month, is made of more than 150k records). Dealing with these large datasets requires efficient processing algorithms. In this context, the proposed place learning algorithm is an interesting contribution since the clustering process is quite efficient.

Acknowledgements. Research group supported by FEDER Funds through the COMPETE and National Funds through FCT – Fundação para a Ciência e a Tecnologia under the Project: FCOMP-01-FEDER-0124-022674.

References

1. Yu, H., Shaw, S.-l.: Representing and Visualizing Travel Diary Data: A Spatio-temporal GIS Approach. In: 2004 ESRI International User Conference, pp. 1–13 (2004)
2. De Almeida, V.T., Güting, R.H., Behr, T.: Querying Moving Objects in SECONDO. In: 7th International Conference on Mobile Data Management MDM 2006, p. 47. IEEE Computer Society (2006)

3. Erwig, M., Güting, R.H., Schneider, M., Vazirgiannis, M.: Abstract and discrete modeling of spatio-temporal data types. In: Proceedings of the Sixth ACM International Symposium on Advances in Geographic Information Systems, GIS 1998, vol. 3, pp. 131–136 (1998)
4. Sevtsuk, A., Ratti, C.: iSPOTS. How Wireless Technology is Changing Life on the MIT Campus. In: Proceedings of the 9th International Conference on Computers in Urban Planning and Urban Management, CUPUM 2005 (2005)
5. Brockmann, D., Theis, F.: Money Circulation, Trackable Items, and the Emergence of Universal Human Mobility Patterns. IEEE Pervasive Computing 7, 28–35 (2008)
6. Reades, J., Calabrese, F., Sevtsuk, A., Ratti, C.: Cellular census: Explorations in urban data collection. IEEE Pervasive Computing 6, 30–38 (2007)
7. Hagen-Zanker, A., Timmermans, H.: A Metric of Compactness of Urban Change Illustrated to 22 European Countries. In: Bernard, L., Friis-Christensen, A., Pundt, H. (eds.). Lecture Notes in Geoinformation and Cartography, pp. 181–200. Springer, Heidelberg (2008)
8. Moreira, A., Santos, M., Wachowicz, M., Orellana, D.: The impact of data quality in the context of pedestrian movement analysis. In: Painho, M., Santos, M., Pundt, H. (eds.) Geospatial Thinking, pp. 61–78. Springer, Heidelberg (2010)
9. Chon, Y., Cha, H.: LifeMap: A Smartphone-Based Context Provider for Location-Based Services. IEEE Pervasive Computing 10, 58–67 (2011)
10. Ester, M., Kriegel, H.-P., Sander, J., Xu, X.: A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. Computer, 226–231 (1996)
11. Ertoz, L., Steinbach, M., Kumar, V.: Finding Clusters of Different Sizes, Shapes, and Densities in Noisy, High Dimensional Data. In: Second SIAM International Conference on Data Mining, pp. 47–58 (2003)