Published in final edited form as:
Trans GIS. 2016 December ; 20(6): 869–886. doi:10.1111/tgis.12181.

Classifying Human Activity Patterns from Smartphone Collected GPS data: a Fuzzy Classification and Aggregation Approach

Neng Wan, Ph.D.[1,*] and **Ge Lin, Ph.D.**[2,*]

[1]University of Utah, Department of Geography, 260 S. Central Campus Dr. Rm 270, Salt Lake City, UT 84112-9155

[2]University of Nevada - Las Vegas, School of Community Health Sciences, Las Vegas, NV 89154

Abstract

Smartphones have emerged as a promising type of equipment for monitoring human activities in environmental health studies. However, degraded location accuracy and inconsistency of smartphone-measured GPS data have limited its effectiveness for classifying human activity patterns. This study proposes a fuzzy classification scheme for differentiating human activity patterns from smartphone-collected GPS data. Specifically, a fuzzy logic reasoning was adopted to overcome the influence of location uncertainty by estimating the probability of different activity types for single GPS points. Based on that approach, a segment aggregation method was developed to infer activity patterns, while adjusting for uncertainties of point attributes. Validations of the proposed methods were carried out based on a convenient sample of three subjects with different types of smartphones. The results indicate desirable accuracy (e.g., up to 96% in activity identification) with use of this method. Two examples were provided in the appendix to illustrate how the proposed methods could be applied in environmental health studies. Researchers could tailor this scheme to fit a variety of research topics.

Keywords

Fuzzy logic; smartphone; GPS; GIS; environmental health

1. Introduction

Human activities are behaviors that individuals take to fulfill specific purposes during their everyday life. Location information about human activities is vital for a variety of environmental health studies such as activity pattern identification (Liao et al. 2007, Wan and Lin 2013), environmental exposure assessment (Elgethun et al. 2003, Wheeler et al. 2010, Wu et al. 2011, Almanza et al. 2012), and clinical interventions (Schenk et al. 2011, Hermann 2011). GPS units could meet most such studies' location measurement requirements because they can record spatial-temporal information more accurately and systematically than traditional measurements such as home address and street intersection.

[*]corresponding authors: Neng Wan, Ph.D. Department of Geography, University of Utah, 260 S. Central Campus Dr., Salt Lake City, UT 84112-9155. 801-585-3972 (office), 801-581-8219 (FAX), 512-757-0309 (cell), neng.wan@utah.edu, Ge Lin, Ph.D., School of Community Health Sciences, University of Nevada - Las Vegas, Las Vegas, NV 89154, ge.kan@unlv.edu.

GPS units' ability to continuously record an individual's location also makes it possible to obtain more complete information than can be obtained with traditional measurements of human activities, such as duration, activity type, trip purpose, and environmental contexts.

More recently, portable GPS units, especially smartphones, have emerged as an alternative equipment for capturing locations to analyze health behaviors and environmental exposure (Stopher 2009, Montoliu and Gatica-Perez 2010, Wiehe et al. 2010, Kerr et al. 2011, Schenk et al. 2011, Wan et al. 2013, Wan and Lin 2013). Compared to traditional GPS units, smartphones not only capture locations from more sources (e.g., GPS, cell tower triangulation, WiFi network locations), but also provide functions such as real-time data transfer, communication, and acceleration monitoring, which are not available from traditional GPS units (Schenk et al. 2011, Carson et al. 2013). In addition, since most individuals have their own smartphone nowadays (Pew Internet Research 2015), this data collection will reduce enrollment costs and greatly alleviate the burden of carrying extra devices. These advantages point to great potential of smartphones in environmental health studies.

However, smartphone-measured GPS data have two limitations with regard to inferring location-based activities. First, smartphone measurements may not be as accurate as those of larger GPS units (Wan et al. 2013). The decreased location accuracy may make the point characteristics that are critical for subsequent classification unreliable (Wan and Lin 2013). Second, the limited battery power of smartphones makes data loss very common during data collection, which poses challenges for further processing of these data. These limitations increase the difficulty in extracting health-related activity information from smartphone-measured GPS traces. Current protocols of GPS data processing were primarily designed for transportation mode detection (Gonzalez et al. 2008, Zheng et al. 2008, Zhang et al. 2011, Gong et al. 2012), travel behavior analyses (Wolf et al. 2001, Auld et al. 2009, Papinski et al. 2009, Schüssler and Axhausen, 2009, Clark and Doherty 2010, Oliver et al. 2010, Millward and Spinney 2011, Millward et al. 2013), and movement analyses (Ashbrook and Starner, 2003, Vazquez-Prokopec et al. 2009). A common practice of these protocols is to divide the sequence of GPS points into segments based on similarities of point characteristics such as speed and acceleration. Then, GIS layers (e.g., street network, bus routes, land use and land cover) and short GPS-promoted surveys to derive such information of individuals (Lee-Gosselin et al. 2006, Gonzales 2008, Gong et al. 2012, Auld 2009, Oliver 2010). For such studies, GPS devices tend to have a fine spatial accuracy by default and the data collection is generally implemented in outdoor situations that are characterized by decent GPS accuracy. Therefore, location inaccuracy was a minor concern for these segment-based methods, and the influence of GPS signal loss on study results was seldom accounted for (Wan and Lin 2013). To extend segment-based GPS data processing to incorporate location specificity, one has to account for location uncertainties in smartphone GPS data processing.

This article presents a classification method that can automatically identify individuals' activity statuses from smartphone GPS data. Taking advantage of fuzzy logic theory, this method could overcome the influence of inaccurate location observation while accommodating the non-linear relationship between point characteristics and activity statuses. In addition, this method draws on the characteristic of point groups rather than the attributes of individual points when classifying GPS trajectories. The proposed scheme will be widely applicable to studies that use popular GPS units such as smartphones to monitor human activities.

Methodological challenges and potential solutions to activity-based GPS data processing

To capture an individual's physical activities and the corresponding environmental factors effectively, GPS data should be processed to infer information such as activity location, type, duration, and environmental contexts (e.g., indoor/outdoor, exposure to pollutants, or proximity to green space). Current methods of GPS data processing are generally segmentation-based (Ashbrook and Starner, 2003, Liao et al. 2007, Guc et al. 2008, Zheng et al. 2008, Schüssler and Axhausen, 2009, Zhang et al. 2011), and are used mainly to identify transportation mode and travel behaviors. For example, a travel behavior study may be interested in when a commuter's activity changes from walking to a bus station to riding on a bus. It is natural using current methods to divide the whole GPS trajectory into segments by characteristics of individual points (e.g., speed, acceleration). It is critical in these methods to determine change points (or critical points) at which an individual's behavior pattern changes (e.g., from staying to walking, or driving to staying). Then, rule-based or deterministic methods can be adopted to determine an individual's status for each segment. Although this scheme has been quite effective in transportation studies, it may not be suitable for environmental health studies for two reasons: 1) it was not designed to account for activity-specific information such as duration and exact location, and 2) GPS units in environmental health studies are prone to signal loss and wide location uncertainties. Therefore, using traditional segmentation-based GPS data processing methods in location-duration, activity-based studies presents several methodological challenges.

First, it is necessary to improve the accuracy of location-based activity inference. Rule-based inference cannot fully capture the non-linear nature of the relationship between point characteristics and activity patterns. For example, there are no strict cut-off values for identifying walking activities because they may differ among people. Using a speed criterion, a stay point with a large deviation from the actual location could easily be misclassified as a traveling or walking point due to the overestimation of speed (Cimon and Wisdom 2004, Wan and Lin 2013). This misclassification could permeate subsequent segmentation process. Therefore, identifying location-based activities requires spatial inference beyond deterministic or rule-based inference, which cannot capture the complex nature of the point-activity relationship. In their methodological development of GPS measures, Kim et al. (2012) combined use of a diary and GPS to determine indoor and outdoor activities. It is suggested that an individual's activity patterns should be validated before tailoring individual's cut-off values. Such a machine-learning process for an 'individual tune-up' may not be necessary for transportation mode detection, because a stationary point would suggest a bus stop or parking lot stay. Given the degraded location accuracy and other uncertainties associated with smartphones, rule-based or linear-nature decision rules may not be appropriate.

Second, although traditional segmentation-based GPS data processing often identifies home and work locations, it is essentially non-location-specific data processing. Each point of interest is used for segmentation but not for location-based activity inference. For activity-based GPS data processing, each activity is accompanied by its spatial stamps that may have different implications for clinical and public health interventions. For instance, in a GPS trajectory that includes an individual traveling to and playing in a park, life-space researchers are generally interested in how far away the park activity is from the home location (Schenk et al. 2011), which is an important indicator of cognitive and physical capacities (Tinetti et al. 1990, Barnes et al. 2007). For the same GPS trajectory, traditional segmentation-based GPS data processing may be interested in only three segments: walking, driving, and being stationary, without the need to distinguish activities in two different locations. In short, GPS data processing in a duration-location study may be based on the same GPS data as in a traditional transportation study, but the purposes of data processing in the two studies differ.

Finally, as implied in the second challenge, environmental health studies and transportation studies have different processing priorities for GPS data. In transportation mode detection, for example, once a segment is determined through both ends of change points, points within a segment become trivial. In activity or exposure classification, however, both segments and containing points within a segment are considered important. In fact, few researchers have considered the homogeneity of neighboring points within a segment, which could help overcome uncertainties from single points. For example, a ‘jumping’ point within a staying activity could be misclassified into driving or walking, but the influence of other staying points around the jumping point could help correct this problem.

It is clear that the non-linear nature of the relationship between smartphone-collected GPS data and an activity of interest requires new, nondeterministic methods. Location information should be accompanied with higher level information about activity. In this paper, we primarily address the nonlinear nature of the relationship between smartphone-collected GPS data and activities of interest. In particular, we employ a fuzzy logic method along with a point aggregation strategy to determine spatial patterns of individual activities.

Fuzzy Logic Methods

Fuzzy logic is a type of reasoning theory that deals with approximate or ambiguous values rather than crispy data. Compared to traditional logic that results in Boolean sets, fuzzy logic methods result in partially true values ranging from 0 (i.e., completely false) to 1 (i.e., completely true). This characteristic makes it possible to deal with vague, inaccurate variables and to capture non-linear relationships, which work well in processing unreliable GPS points collected by smartphones.

In general, a fuzzy logic method can be implemented in three steps (shown in Figure 1): input fuzzification, fuzzy inference, and output calculation. The input fuzzification process transforms crispy input values into linguistic terms using a membership function (MF) that projects an input value into a membership degree between 0 and 1. Suppose $X = \{x_1, x_2, \ldots\ldots, x_m\}$ represents a vector of input variables, and x_i (i=1,…m) denotes the ith variable. An MF maps each x_i into the probabilities of belonging to different linguistic terms or fuzzy sets $T(x_i)=\{T^1_{x_i}, T^2_{x_i}, \ldots\ldots, T^p_{x_i}\}$. For example, if the first input variable, x_1, represents the speed of a GPS point, then the corresponding descriptive terms, $T(x_1)=\{T^1_{x_1}, T^2_{x_1}, \ldots\ldots, T^p_{x_1}\}$, might be *fast*, *slow*, and *zero*. In this case, *p* is 3 and the MF derives the probabilities of this speed being each of the three terms. Based on these probabilities, the fuzzy inference engine uses a set of *if-then* rules to specify the non-linear relationship between input fuzzy sets and output fuzzy sets. In this example, if the outcome of interest is the probability of driving, then a rule might be "if the speed is *fast* then the probability of *driving* is *high*". And the output fuzzy set (or output linguistic variables) $T(y)=\{T^1_y, T^2_y, \ldots\ldots, T^q_y\}$ might be *very low*, *low*, *medium*, *high,* and *very high*. In this case, *q* equals 5. The inference results of all rules are then processed to derive crispy output values. Note that this is an example based on a single input variable (i.e., speed). In case of multiple input variables, similar procedures could be implemented for each variable and their joint influence on the output linguistic variables would be assessed during the output calculation.

Depending on how the crispy output values are calculated, fuzzy logic methods can be categorized into two major types: Mamdani type (Mamdani and Assillian 1975) and Sugeno type (Takagi and Sugeno 1985). The former uses a defuzzification process from the output fuzzy set to generate the crispy output values according to different aggregation techniques. The latter uses a weighted average of the output fuzzy set to represent the output values. Sugeno-based fuzzy method is considered better than Mamdani's method because it is more flexible to realize and more computationally efficient (Jassbi et al. 2006, Piolet 2006, Shleeg and Ellabib 2013).

2. Methodological Development of Fuzzy Classification

To derive detailed information about human activities from GPS data, we developed a three-step fuzzy logic classification scheme. The first step is to select key input variables that could effectively distinguish different activity patterns. The second step is to apply the fuzzy theory to estimate the probability of human activity status. The final step is to derive activity patterns through point aggregation and smoothing. The first two steps directly benefit from the fuzzy logic theory, and the final step overcomes potential limitations of fuzzy methods such as misclassification.

2.1 GPS-measured Activities

Considering the characteristics of GPS points, this study categorizes an individual's activity statuses into three groups: staying, walking, and other transportation. Staying activities are those when an individual stays around a location (e.g., shopping, dinning out); walking activities refer to outdoor walking; and other transportation refers to motorized or non-motorized transportation (e.g., car driving, bus riding, and cycling) other than walking. We emphasize outdoor for the walking status because when an individual walks inside a building, their GPS points may exhibit a clustering pattern around the building location due to satellite signal shielding. In this case, the location deviation of GPS points makes it difficult to distinguish whether the individual was walking or not. The three statuses cover almost all types of everyday activities, which makes them suitable for representing continuous activities of individuals. In addition, these statuses provide the foundation for deriving more detailed activity statuses. For example, other transportation may be further classified into cycling, driving, or bus riding based on GPS-derived acceleration or other types of sensor data (Gonzales et al. 2008; Oliver et al. 2010). Based on the activity statuses, researchers could derive more detailed activity information such as duration, frequency, type, variability, purpose, etc.

2.2 Selecting input variables

After exploratory analyses of common point characteristics, we selected point speed and point angle as the input variables. Speed is the most commonly used criterion in GPS point classification, as different activity patterns tend to have different speed markers. In general, the speed of a GPS point can be calculated by dividing the cardinal distance between it and its previous point by the difference between their time stamps. However, GPS signal losses from factors such as shielding and battery exhaustion could heavily influence this simple calculation (Wan and Lin 2013). To amend this effect, it is necessary to adopt some location-distance criteria in speed calculation. Wan and Lin (2013) reported that a long distance or time gap between a point and the previous point often suggests an anomaly spike, and therefore, should be treated. Following their practice, the speed of the point is set to either zero or that of the previous point.

However, using speed alone is not sufficient for differentiating human activity status (Zheng et al. 2008). Point angle, which can be defined from a vantage point in reference to its previous and succeeding points, provides another important variable for activity classification. Therefore, we introduced a point angle indicator to distinguish activity patterns from GPS points. It is noted that point angles tend to be large when a subject is moving (e.g., walking, other transportation), and small when a subject is staying. This characteristic is shown in Figure 2: (a) defines an angle θ_i for a point P_i, (b) signals a moving trajectory when the contiguous angles are large, and (c) suggests a staying pattern. It is clear that angle points are effective in distinguishing staying and moving points.

GPS-measured speed and direction have been proved effective in measuring human activity patterns (Zheng et al. 2008). Our assessment suggested that the two criteria above are sufficient to distinguish activity patterns of staying, walking, and other transportation. Speed provides a general activity profile, while angle complements it by accounting for errors due to location inaccuracy. These two variables are therefore selected as the input for the fuzzy classification. The input vector is denoted as $X = \{x_1, x_2\}$, where x_1 is point speed and x_2 is point angle.

2.3 Implementing fuzzy classification

Based on input values of point speed and point angle, we can generate preliminary probabilities of a point belonging to either staying, walking, or other transportation. The fuzzy set associated with point speed is denoted as $T(x_1)=\{T_{x_1}^1, T_{x_1}^2, T_{x_1}^3\}$, which includes three linguistic variables: *zero*, *slow*, and *fast*. The fuzzy set associated with point angle is denoted as $T(x_2)=\{T_{x_2}^1, T_{x_2}^2\}$, which includes two linguistic variables: *narrow* and *wide*. Point speed and point angle are fuzzified using membership functions shown in Figure 3, where the Y-axis scales membership probabilities. We used Z-, trapezoidal-, and S-shaped membership functions to represent the speed of zero, slow, and fast, respectively. As the speed moves from zero to fast (Figure 3a), the probabilities of belonging to each membership state also changes. The angle characteristics of narrow and wide are denoted using Z-shaped and S-shaped membership functions, respectively. Likewise, as the angle changes from below 90 degrees to above 90 degrees (Figure 3b), the probabilities of belonging to moving and staying also change. Since the speed of zero, slow, and fast was primarily designed to distinguish staying, walking, and other transportation, the ranges for the three speeds were determined for this purpose. For example, the speed of 2.78 m/s has been widely acknowledged as the threshold for differentiating transportation and non-transportation statuses (Schüssler and Axhausen 2009). Therefore, the breaking value between slow speed (which primarily corresponds to walking) and fast speed (which primarily corresponds to other transportation) was set to be around 2.78 m/s. In addition, 1.25m/s, which has been suggested as the speed that most young adults could manage (Knoblauch et al. 1996), was set to represent the lower bound of slow speed. We adopted this value because the evaluation data in this study were collected by younger adults (i.e., <65). For point angle, we set the breaking value between narrow and wide to be 90 degrees because few turns (e.g., at a road intersection) during other transportation or walking activities have an angle smaller than this value.

Since our purpose is to determine whether a point belongs to walking, staying, or other transportation, we adopted three inference systems, each corresponding to one outcome status. For each outcome, the output fuzzy set is denoted as $T(y)=\{T_y^1, T_y^2, T_y^3\}$, which includes three linguistic variables: *zero*, *low*, and *high*. Table 1 reveals our conceptual assumptions for the fuzzy rules. As shown in the table, zero speed with narrow angle would suggest high probability of staying; fast speed with wide angle would suggest high probability of other transportation; and slow speed with wide angle would suggest high probability of walking. Other rules were determined based on empirical experiences and common sense. For example, a zero speed with wide angle may indicate either a brief stop during a walking activity or an indoor staying activity. Therefore, the probability of staying and walking was both set to be medium; a fast speed with narrow angle may be due to the great deviation error of an indoor GPS point or the turnings during a driving activity. The probability for the two statuses was both set to be medium. When implementing the rules in our method, the probability value for the three fuzzy outcomes (i.e., zero, low, high) was set to be 0, 50%, and 100%, respectively. The result output step adopted the Sugeno-type model due to its two major advantages mentioned in the last section.

2.4 Point segmentation and aggregation

After fuzzy inference has estimated the probabilities of staying, walking, or other transportation for each point, it is necessary to aggregate points into segments so that activity-specific information can be derived. However, due to misclassifications and data noise, the initial probabilities of points could not be used directly for segmentation. Therefore, we introduced a point smoothing and aggregation process, which uses the status of neighboring points to 1) overcome individual point uncertainty and 2) to derive meaningful information about an individual's activities.

First, the probabilities generated by the fuzzy classification were smoothed using a Gaussian kernel. Specifically, we select the two previous points and the two subsequent points within 120 seconds of the current point. If there are fewer than 2 points within the period, only those within the period are selected for smoothing. We adopted two neighbors and 120 seconds because neighboring points that are too far away (either in sequence or in time range) have much higher likelihood to be with a different activity status than the central point. The smoothed probability is calculated as

$$SP_i=\frac{\sum P_{ij}W_j}{\sum W_j} \quad (1)$$

where SP_i is the smoothed probability of the current point for the ith status (i=1,2,3), P_{ij} is the original probability of the ith status for the jth neighbor, and W_j represents the Gaussian weight for the jth neighbor. The Gaussian weights for the five points were set to be W = [0.1, 0.3, 1, 0.3, 0.1]. We used an evaluation strategy (Wan et al. 2012) to compare this group of weights with other groups that have different decay effects (e.g., [0.2, 0.3, 1, 0.3, 0.2]), and the results indicate that this group yields better classification quality (results not shown here). The preliminary status (e.g., staying, walking, or other transportation) of a point is determined as the one with the highest smoothed probability.

The second step is to aggregate individual GPS points into segments. Based on the smoothed probabilities, aggregating points into segments is mostly straightforward. However, some noise still remains. For example, during an ‘other transportation’ (e.g., driving) activity, an individual may occasionally stop in traffic; GPS points observed when a driver or walker turns around may exhibit as staying points. These points may divide the other transportation activity or the walking activity into several shorter ones. While the “stop breaks” may be useful for transportation studies, they are generally not of value for environmental health topics such as air pollution exposure assessment. We use a seed-growing strategy to remove these lingering uncertainties. Specifically, seed segments are first identified based on the preliminary status and a minimum point number threshold (NT). For example, if a group of consecutive points (n>NT) have the same status (e.g., other transportation), then this group of points is identified as a seed segment with the same status of the points. We set NT to four because the likelihood that four consecutive points are mistakenly classified as the same status is very low. A larger NT (i.e., NT>4) poses a higher requirement for the continuity of segments and may miss potential activities. Note that due to the aforementioned uncertainties, consecutive seed segments are not necessarily connected, because some points between them do not belong to either segment (shown in Figure 4).

To determine if an in-between point belongs to the previous or the subsequent segment, we designed a set of rules to aggregate these points:

1. If the previous and subsequent seed segments have the same attribute, then all points between them are assigned the same attribute.
2. If the two seed segments have different statuses, then the accumulated probabilities of the two statuses among all connecting points are calculated, and

the status with the larger accumulated probability is selected for all the connecting points.

In addition to the seed segment identification and segment aggregation described above, we also introduced a distance criterion to distinguish segments: any point with a distance longer than 200 meters from its previous point is set to start a new segment, regardless of the similarity of point attributes. This criterion was set to adjust for the influence of GPS signal loss, as a long distance gap may represent the beginning of another activity, regardless whether these two activities have the same attribute. We use 200 meters because it represents the longest diagonal length of buildings in this study. Building shielding is the most frequent reason for GPS signal loss.

3. Methodology Evaluation

3.1 Data

To evaluate the proposed method, we enrolled two mid-aged, male subjects to collect continuous GPS data for two weeks during the period from December 2012 to March 2013. Specifically, one subject carried a Samsung Galaxy Note I phone and the other carried a Motorola Droid phone; both were running on the Android OS. An app was set to automatically collect GPS points at a 5-second interval in the background of the OS. The subjects were asked to power on the phone when they got out of bed in the morning, carry the phone with them as much as possible, and charge the battery before they went to bed at night. During the two-week observation period, both subjects were also asked to use a paper journal to record real-time information about their everyday activities, such as start time, end time, location, trip purpose, approximated speed (for outdoor walking), and shielding conditions (e.g., open, near tree, near wall, highly shielded). After each week's data collection, the subject was asked to correct mistakenly recorded information (e.g., time, location) and make up for activities that the subjects forgot to record but were reflected by the GPS data by overlaying GPS points (with timestamp information) on GIS layers (e.g., streets, building parcels) (Wan and Lin 2013). In addition, two weeks of GPS data were selected from a previous study (Wan and Lin 2013) in which a mid-aged, male subject carried a Nokia N900 smartphone for four months from January 2012 to April 2012. Although using a different OS (i.e., Maemo), the Nokia phone was set to collect GPS data using the same protocol as the current study. All GPS data in the three datasets underwent a cleaning procedure (Wan and Lin 2013) to remove obviously incorrect points. The data collection and analysis have been approved by the IRB of the University of Nebraska Medical Center.

3.2 Evaluation procedures

In the absence of empirical evaluation criteria, we proposed 4 indicators to assess the performance of the proposed method: under detection rate, over detection rate, duration deviation, and location deviation. We first define a baseline activity as one that was recorded in the diary or derived from the GIS-based recall. When a detected activity does not match any baseline activity in terms of starting time, ending time, location, and activity type, it is considered a case of over detection; when a baseline activity was not detected from the GPS points by the method, it is a case of under detection. Based on these assumptions, the over detection rate (ODR) is calculated as the ratio between the number of over detection cases and the number of identified activities, and the under detection rate (UDR) is calculated as the ratio between the number of under detection cases and the number of baseline activities. Note that ODR and UDR are not seamlessly complementary (i.e., ODR+UDR<1) because they have different denominators.

Duration deviation refers to the sum of deviations of the starting time and the ending time between the baseline and the detected activities (if detected). Location deviation refers to the deviation from detected staying activities to the real locations of these activities.

We evaluated location deviation because it is an important factor for inferring trip purposes and environmental exposures (Elgethun et al. 2003, Wheeler et al. 2010, Wu et al. 2011, Almanza et al. 2012). We focus on staying activities when a particular location can be easily calculated and evaluated. A recursive kernel density method (Wan and Lin 2013) was used because it generates more accurate locations for cluster-shaped activities. This method first divides the cover area into 9 even cells, calculates the kernel density of each cell, and then selects the cell with the highest density, which is the candidate for the next round. The cell division repeats until the cell size is smaller than a threshold value (for example, 5 meters) and the location of the activity is determined as the centroid of the last cell with the highest density.

Our evaluation was based the original GPS data set. In addition, we resampled the original GPS data sets into larger intervals and examined the change of classification quality among different intervals. We did this for two reasons. First, GPS logging interval is an important consideration for smartphone data collection, because a very short interval speeds up phone battery depletion and causes substantial loss of data (Krenn et al. 2011, Oliver et al. 2010, Vazquez-Prokopec et al. 2009). It is therefore critical to determine an observational interval that is wide enough to provide comparable levels of activity information while consuming less power. Second, shorter intervals do not necessarily lead to better classification results, because speeds and point angles calculations tend to be unreliable at short intervals (Cimon and Wisdom 2004, Wan and Lin 2013). Therefore, it is necessary to incorporate the influence of observation interval when evaluating the classification method. To do this, we resampled the original 5-second data to larger intervals (up to 100-second) with the increment of 5 seconds. The proposed method was implemented on both the original data and the resampled data. We then compared the four indicators across observation intervals. To describe the overall classification quality for a specific observational interval, we also integrated the four single indicators into a Classification Quality Index (CQI) (Wan and Lin 2013). Specifically, this index assigns a weight to each of the four indicators and summarizes the weighted indicators to denote the overall classification quality. Since over detection and under detection are the major sources of classification error, the weight for ODR and UDR was set to be double (i.e., 0.333) of that for duration deviation and location deviation (i.e., 0.167).

3.3 Results

Table 2 shows the overall characteristics of the collected data by phone models. The Samsung and Motorola models collected 9,544 and 29,992 points, respectively. The Samsung phone collected fewer points because the subject who carried this phone was more likely to stay indoors, which leads to frequent signal loss. For the same reason, this subject's phone also collected many fewer reference activities.

Table 3 lists the classification results in terms of the four single indicators as well as the CQI. As shown in Table 3, data based on the original observational interval yields the highest number of detected activities (n=638), with a 15.7% ODR and a 4% UDR. After adjusting for ODR and UDR, the number is still higher than for the reference activities because a number of reference activities were divided into multiple segments leading to over-detected activities. The matching rate was 96% at the 5-second interval. The rate remained high (94%) at the 15-second interval and declined steadily thereafter, leading to an increased chance of missed activities. Although the number of over-detected activities may have been different, the over detection rate remained relatively stable with the increase of the interval. This result is to be expected because increased observation intervals with some loss of information should not generate additional activities. The change patterns of ODR and UDR are shown in Figure 5.

The next two columns of Table 3 list location and duration deviations. Note that the location deviation was assessed for staying activities only. It is a relative accuracy indicator because it was in reference to the shortest observation interval (i.e., 5 seconds). It seems that location deviations increased steadily as the observation intervals increase, while duration deviations oscillated around 4.5 minutes. At the 15-second interval, the location deviation was about 13 meters while the duration deviation was less than 3.5 minutes, and both were reasonably small. A similar pattern could also be observed for CQI (shown in Figure 6). According to the definition, a larger CQI represents better classification quality, and vice versa. As shown in Figure 6, the interval of 15 seconds yields the best CQI. After 15 seconds, the CQI exhibits a decreasing trend as the observation interval increases.

4. Discussion and Conclusion

The increasing popularity of the smartphone has made it a promising tool for measuring human activities in environmental health studies. However, uncertainties about location measurement prevent the effective use of smartphone GPS data. The current study proposes a series of methods to overcome uncertainties from smartphone collected GPS data. Specifically, a fuzzy classification method was developed to account for unstable locations and the non-linear relationship between GPS characteristics and desired output variables. A segmentation method was then proposed to adjust for equipment-irrelevant data noises such as data loss and trivial fragments (e.g., traffic stops). Our idea of integrating fuzzy logic and segmentation provides new perspectives to location uncertainty studies.

In general, fuzzy logic classification requires empirical criteria or cut-off parameter estimates. In this study, the cut-off values for point speed and point angle in membership functions are based on previous studies and empirical experiences, but they may not fit all movement profiles of all population. For example, the threshold speed of 1.25m/s used to distinguish low and zero speed was measured for adults younger than 65. For studies that involve older adults, the value could be slightly lower because older adults generally walk slower than younger adults (Knoblauch 1996). Also, walking speed varies by trip purpose (e.g., recreational, active transport), even for the same person (Millward et al. 2013), which makes it harder to define gold-standard threshold values. Although the fuzzy classification may be influenced by these uncertainties, the subsequent probability smoothing and segment aggregation methods could overcame the sensitivity to cut-off parameters. This means that, although the change in parameter values may cause some fluctuations in the raw probabilities in fuzzy inference, the influence of such fluctuation can be minimized during probability smoothing. Furthermore, the segment aggregation can further correct misclassified points by grouping neighboring points.

Similar to results of previous studies (Cimon and Wisdom 2004, Wan and Lin 2013), our results suggest that the original 5-second intervals and the resampled 10-second intervals yielded lower CQI than longer intervals up to 30 seconds. The degraded classification quality for the shorter intervals may be due to the input variable calculation since speeds and angles tend to be more unstable at shorter intervals. These findings also indicate the importance of assessing the influence of GPS recording frequency on activity classification, since an inappropriately selected interval may yield unsatisfactory results.

Over detection was due largely to random errors of GPS points rather than to processing algorithms. In situations of signal shielding, the GPS points could be very unstable and sometimes exhibit a curve pattern that could easily be classified as a walking activity (shown in Figure 7). Although this problem cannot be resolved by the algorithms alone, other smartphone data sources may provide a solution. For example, the phone accelerometer can record the three-dimensional acceleration of the phone, from which the physical intensity of the subject can be inferred (Doherty 2009, Schenk et al. 2011, Carson et al. 2013). Previous studies indicated that acceleration data, measured by smartphones or separated units, have great potential in distinguishing different activity patterns (e.g., walking, other transportation, and staying), calculating energy consumption, and inferring gait speeds. Such information could be used to complement the GPS-derived activities and to correct falsely detected activities.

In addition to integrating with accelerometer data, some future works are needed to refine this method. First, we need to distinguish more activity patterns within the ‘other transportation’ category. Besides walking, public health researchers are also interested in active transportation types such as running/jogging, cycling and public transportation. The fuzzy logic method, when combined with multiple GPS-derived indicators (e.g., acceleration), could help achieve this goal. Those indicators do not have to be limited to a single GPS point (for speed and acceleration indicators) or three points (for angle based indicators). One could use summary indicators (e.g., mean, median, variance, and maximum) within a moving window (e.g., 9 consecutive points) to distinguish activity statuses within the ‘other transportation’ category. Second, although the proposed scheme achieved 96% accuracy (which is comparable or slightly higher than most studies) in activity identification at the 5-second interval, a comparison between this scheme and existing methods based on a common dataset would help us better understand the strength and limitation of fuzzy logic in GPS data classification. In addition, the algorithms need to be tested on GPS data collected by newer smartphone models which have improved GPS chips and multi-mode (e.g., GPS, WiFi) location measurement.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Research reported in this publication was supported by the National Institute On Aging of the National Institutes of Health under Award Number R03AG044759. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Almanza E, Jerret M, Dunton G, Seto E, Pentz A. A study of community design, greenness, and physical activity in children using satellite, GPS, and accelerometer data. Health & Place. 2012; 18:46–54. [PubMed: 22243906]

Ashbrook D, Starner T. Using GPS to learn significant locations and predict movement across multiple users. Personal and Ubiquitous Computing. 2003; 7:275–286.

Auld J, Williams C, Mohammadian K, Nelson P. An automated GPS-based prompted recall survey with learning algorithms. Transportation Letters: the International Journal of Transportation Research. 2009; 1:59–79.

Barnes L, Bienias L, Mendes de Leon F, Kim N. Correlates of life-space in a volunteer cohort of older adults. Experimental Aging Research. 2007; 33:77–93. [PubMed: 17132565]

Carlson RH, Huebner DR, Hoarty CA, Whittington J, Haynatzki G, Balas MC, Schenk AK, Goulding EH, Potter JF, Bonasera SJ. Treadmill gait speeds correlate with physical activity counts measured by cell phone accelerometers. Gait & Posture. 2012; 36:241–248. [PubMed: 22475727]

Cimon, N., Wisdom, M. ACCURATE VELOCITY ESTIMATES FROM INACCURATE GPS DATA; Proceedings of the Tenth Forest Service Remote Sensing Applications Conference; April 5–9 2004; Salt Lake City, Utah. 2004.

Clark AF, Doherty ST. A multi-instrumented approach to observing the activity rescheduling decision process. Transportation. 2010; 37:165–181.

Doherty ST. Emerging methods and technologies for tracking physical activity in the built environment. Transport Survey Methods: Keeping up with a Changing World. 2009; (2009):153–190.

Elgethun K, Fenske RA, Yost MG, Palcisko GJ. Time-location analysis for exposure assessment studies of children using a novel global positioning system instrument. Environmental Health Perspectives. 2003; 111:115–122. [PubMed: 12515689]

Gong HM, Chen S, Bialostozky E, Lawson CT. A GPS/GIS method for travel mode detection in NewYork. Computers, Environment and Urban Systems. 2012; 36:131–139.

Gonzalez, P., Weinstein, J., Barbeau, S., Labrador, M., Winters, P., Georggi, N., Perez, R. Automating mode detection using neural networks and assisted GPS data collected using GPS-enabled mobile phones; 15th World Congress on Intelligent Transportation Systems; November 16–20 2008; New York, New York. 2008.

Guc, B., May, M., Saygin, Y., Körner, C. Semantic annotation of GPS trajectories; 11th AGILE International Conference on Geographic Information Science; Girona, Spain. 2008.

Herrmann SD, Snook EM, Kang M, Scott CB, Mack MG, Dompier TP, Ragan BG. Development and validation of a movement and activity in physical space score as a functional outcome measure. Archives of Physical Medicine and Rehabilitation. 2011; 192:1652–1658.

Jassbi, J., Serra, P., Ribeiro, RA., Donati, A. Comparison of Mamdani and Sugeno Fuzzy Inference Systems for a Space Fault Detection Application; Proceeding of the 2006 World Automation Congress (WAG 2006); 2006.

Knoblauch RL, Pietrucha MT, Nitzburg M. Field studies of pedestrian walking speed and start-up time. Transportation Research Records. 1996; 1538:27–38.

Kerr J, Duncan S, Schipperjin J. Using global positioning systems in health research: a practical approach to data collection and processing. American journal of preventive medicine. 2011; 41:532–540. [PubMed: 22011426]

Krenn PJ, Mag DI, Titze S, Oja P, Jones A, Ogilvie D. Use of global positioning systems to study physical activity and the environment. American Journal of Preventive Medicine. 2011; 41:508–515. [PubMed: 22011423]

Lee-Gosselin Martin EH, Doherty ST, Papinski D. Internet-based prompted recall diary with automated gps activity-trip detection: System design. Transportation Research Board 85th Annual Meeting. 2006 No. 06-1934. 2006.

Liao L, Fox D, Kautz H. Extracting places and activities from GPS traces using hierarchical conditional random fields. International Journal of Robotics Research. 2007; 26:119–134.

Mamdani EH, Assilian S. An experiment in linguistic synthesis with a fuzzy logic controller. International Journal of Man-Machine Studies. 1975; 7:1–13.

Millward H, Spinney J. Time use, travel behavior, and the rural-urban continuum: results from the Halifax STAR project. Journal of Transport Geography. 2011; 19:51–58.

Millward H, Spinney J, Scott D. Active-transport walking behavior: destinations, durations, distances. Journal of Transport Geography. 2013; 28:101–110.

Montaliu, R., Gatica-Perez, D. Discovering human places of interest from multimodal mobile phone data; Proceedings of 9th International Conference on Mobile and Ubiquitous Multimedia, MUM; 2010.

Oliver M, Badland H, Mavoa S, Duncan MJ, Duncan S. Combining GPS, GIS and accelerometry: methodological issues in the assessment of location and intensity of travel behaviours. Journal of Physical Activity and Health. 2010; 7:102–108. [PubMed: 20231761]

Papinski D, Scott DM, Doherty ST. Exploring the route choice decision-making process: A comparison of planned and observed routes obtained using person-based GPS. Transportation research part F: traffic psychology and behaviour. 2009; 12:347–358.

Pew Internet Research. U.S Smartphone Use in 2015. 2015 [last accessed 2015/06/19] (http://www.pewinternet.org/2015/04/01/us-smartphone-use-in-2015/).

Piolet Y. Comparison of Mamdani and Sugeno Fuzzy Inference Systems. Work paper. 2006 [last accessed 2014/04/13] (http://www.egr.msu.edu/~thulasid/08IT6018_mamdani_sugeno.pdf).

Schenk AK, Witbrodt BC, Hoarty CA, Carlson RH, Goulding EH, Potter JF, Bonasera SJ. Cellular telephones measure activity and life-space in community-dwelling adults: proof of principle. Journal of the American Geriatric Society. 2011; 59:345–352.

Schüssler, N., Axhausen, KW. Processing GPS raw data without additional information; Paper presented at the 88th Annual Meeting of the Transportation Research Board; January 2009; Washington, DC. 2009.

Shleeg A, Ellabib I. Comparison of Mamdani and Sugeno Fuzzy Inference Systems for the Breast Cancer Risk. International Journal of Computer, Information Science and Engineering. 2013; 7:83–87.

Stopher PR. Collecting and processing data from mobile technologies. Transport Survey Methods: Keeping Up with a Changing World. 2009:361–391.

Tinetti ME, Ginter SF. The nursing home life-space diameter. A measure of extent and frequency of mobility among nursing home residents. J Am Geriatr Soc. 1990; 38:1311–1315. [PubMed: 2254569]

Takagi K, Sugeno M. Fuzzy identification of systems and its applications to modeling and control. IEEE Trans on Systems, Man, and Cybernetics. 1985; 15:116–132.

Vazquez-Prokopec G, Stoddard ST, Paz-Soldan V, Morrison AC, Elder JP, Kochel TJ, Scott TW, Kitron U. Usefulness of commercially available GPS data-loggers for tracking human movement and exposure to dengue virus. International Journal of Health Geographics. 2009; 8:e68.

Wan N, Lin G. Life-space characterization from cellular telephone collected GPS data. Computers, Environment and Urban Systems. 2013; 39:63–70.

Wan N, Qu W, Whitington J, Witbrodt B, Henderson M, Goulding E, Schenk A, Bonasera S, Lin G. Assessing smart phones for generating life space indicators. Environment and Planning B. 2013; 40:350–361.

Wan N, Zhan FB, Zou B, Chow T. A relative spatial access assessment approach for analyzing potential spatial access to colorectal cancer services in Texas. Applied Geography. 2012; 32:291–299.

Wheeler B, Cooper A, Page A, Jago R. Greenspace and children's physical activity: a GPS/GIS analysis of the PEACH project. Preventive Medicine. 2010; 51:148–152. [PubMed: 20542493]

Wiehe SE, Carroll AE, Liu GC, Haberkorn KL, Hoch SC, Wilson JS, Fortenberry JD. Using GPS-enabled cell phones to track the travel patterns of adolescents. International Journal of Health Geographics. 2008; 7

Wolf J, Randall G, William B. Elimination of the travel diary: Experiment to derive trip purpose from global positioning system travel data. Transportation Research Record: Journal of the Transportation Research Board. 2001; 1768:125–134.

Wu J, Jiang C, Houston D, Baker D, Delfino R. Automated time activity classification based on global positioning system (GPS) tracking data. Environmental Health. 2011; 10:101. [PubMed: 22082316]

Zhang L, Dalyot S, Eggert D, Sester M. Multi-stage approach to travel-mode segmentation and classification analysis of GPS traces. ISPRS Workshop on Geospatial Data Infrastructure: from data acquisition and updating to smarter services. 2011:87–93.

Zheng, Y., Liu, L., Wang, L., Xie, X. Learning transportation mode from raw GPS data for geographic applications on the web; Paper presented at the 17th World Wide Web conference; April 2008; Bejing. 2008.

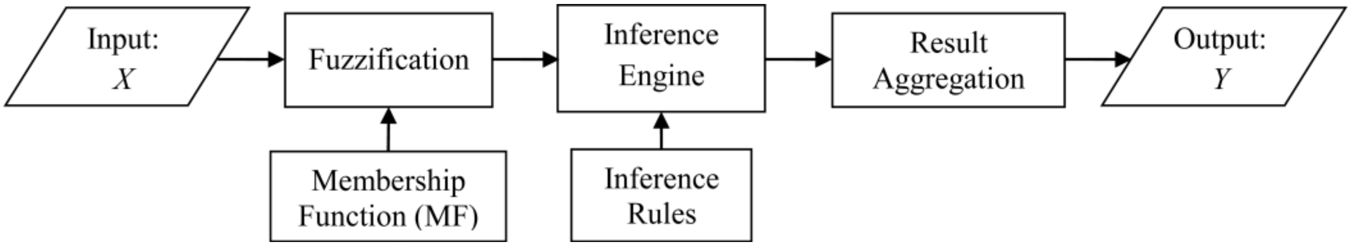

Figure 1.
General flow of fuzzy logic methods

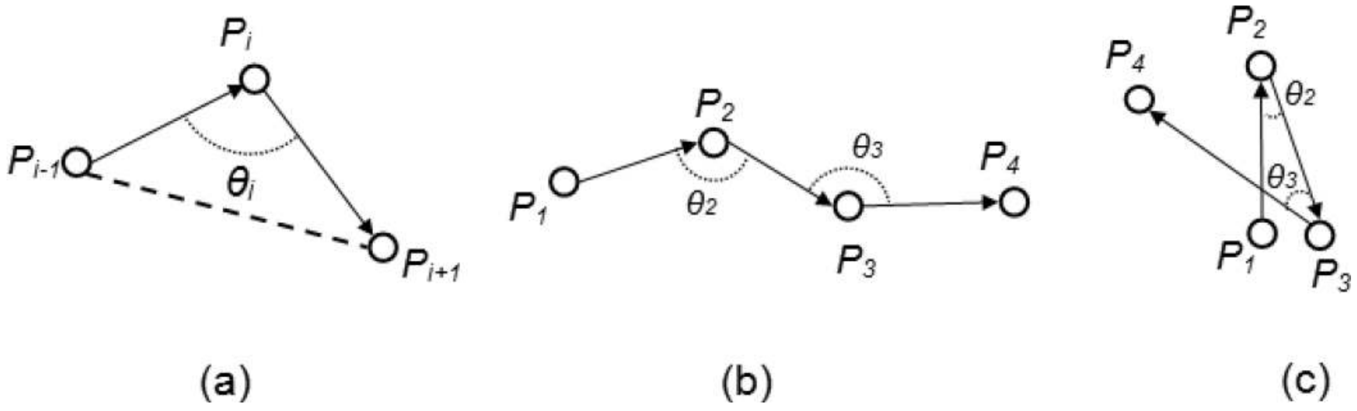

Figure 2.

Definition and characteristics of point angle. (a). Definition of point angle; (b) Point angle of walking points; (c). Point angle of staying points. Due to the location error, GPS points of indoor and outdoor staying activities exhibit a ‘clustering’ pattern, which leads to narrower point angles than those of walking or other transportation.

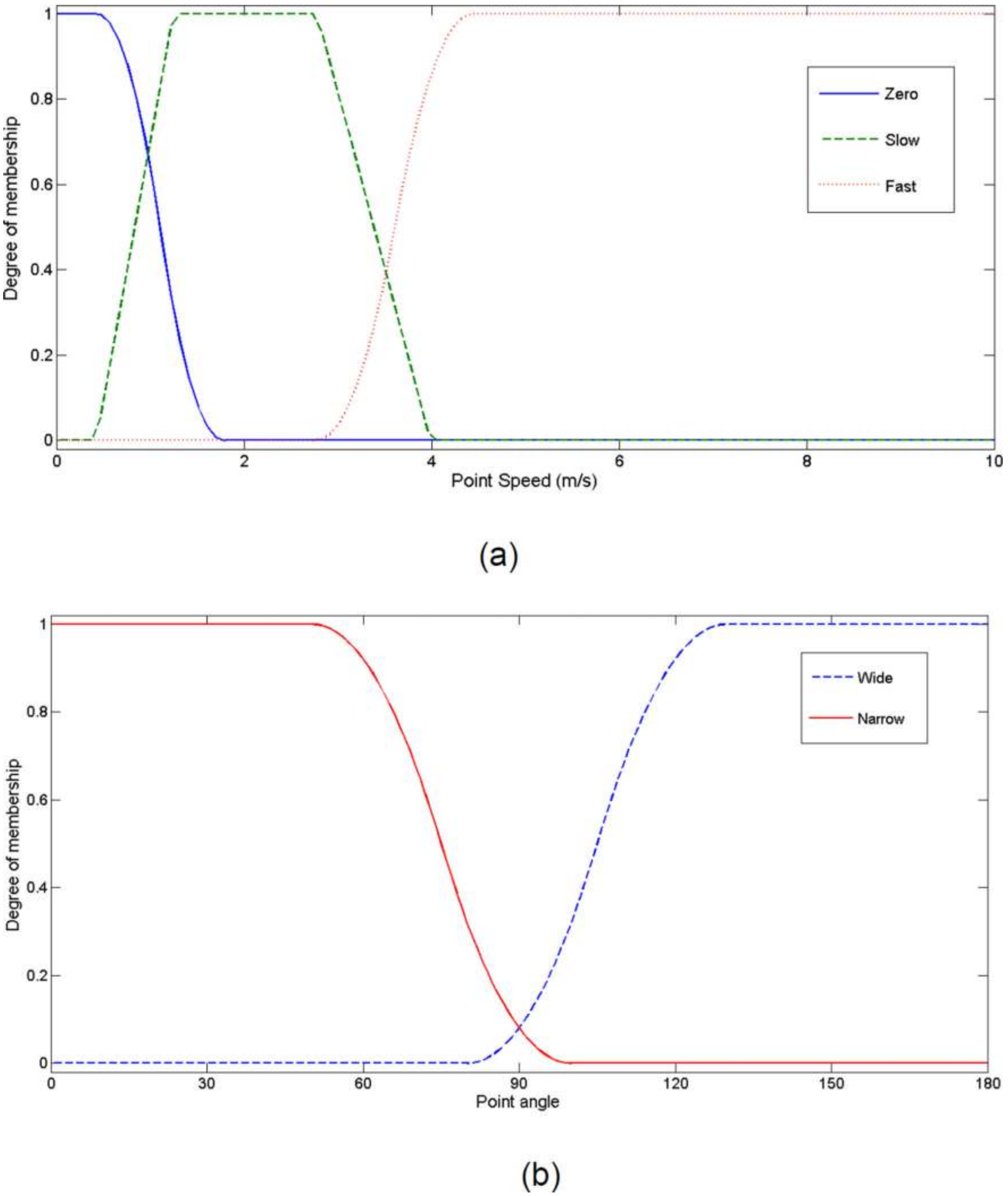

Figure 3.
Membership functions for point speed and point angle

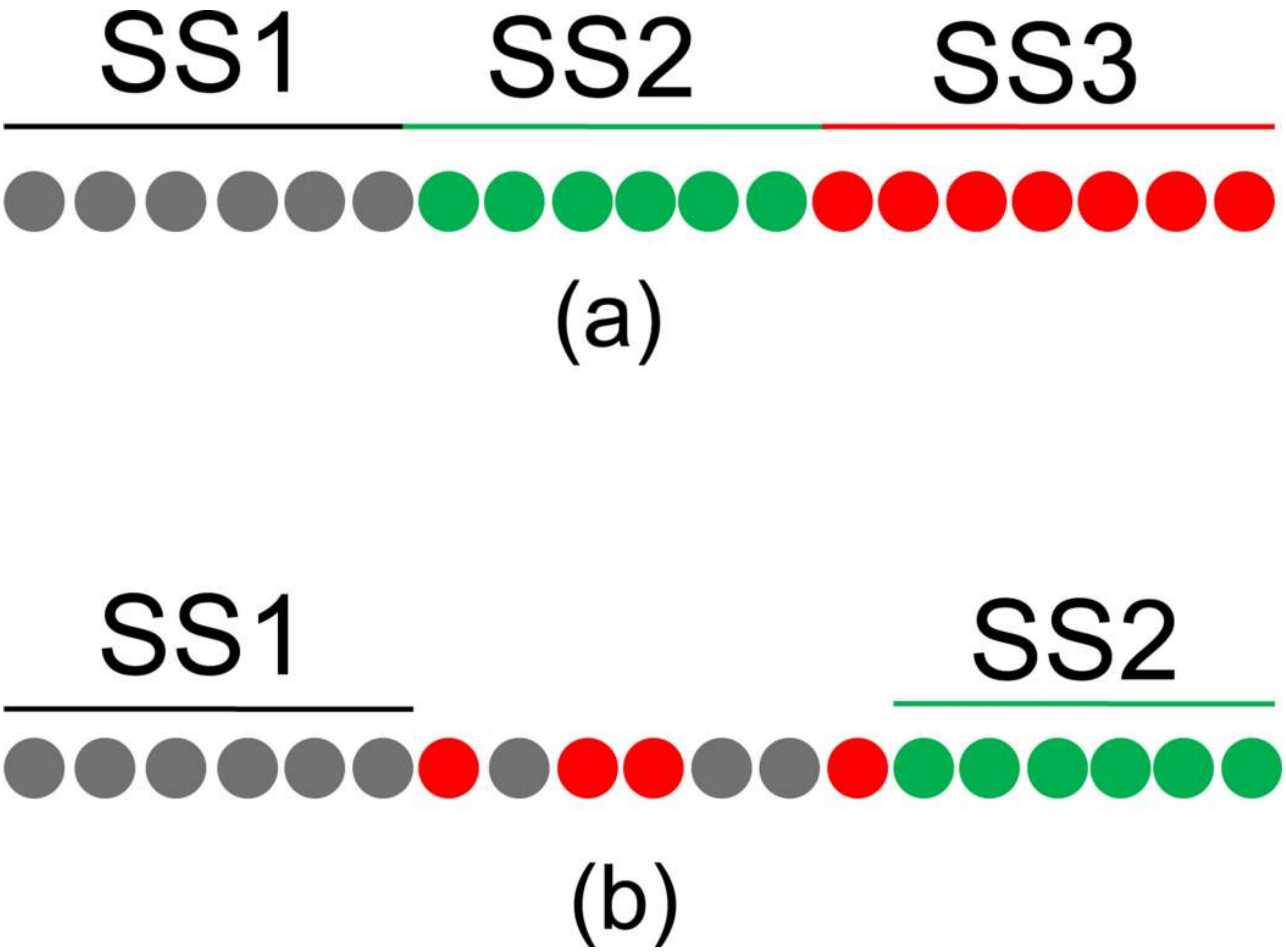

Figure 4.

Illustrations of seed segments (SS), in which a grey point indicates staying, a green point indicates walking, and a red point indicates other transportation. (a). Perfect conditions where seed segments (e.g., SS1, SS2, and SS3) are connected with each other. (b) Conditions where seed segments (e.g., SS1, SS2) are not connected; the seven points between SS1 and SS2 did not form a segment and need to be re-assigned. Note that the location information of points was removed so that points are ranked only by their temporal sequence.

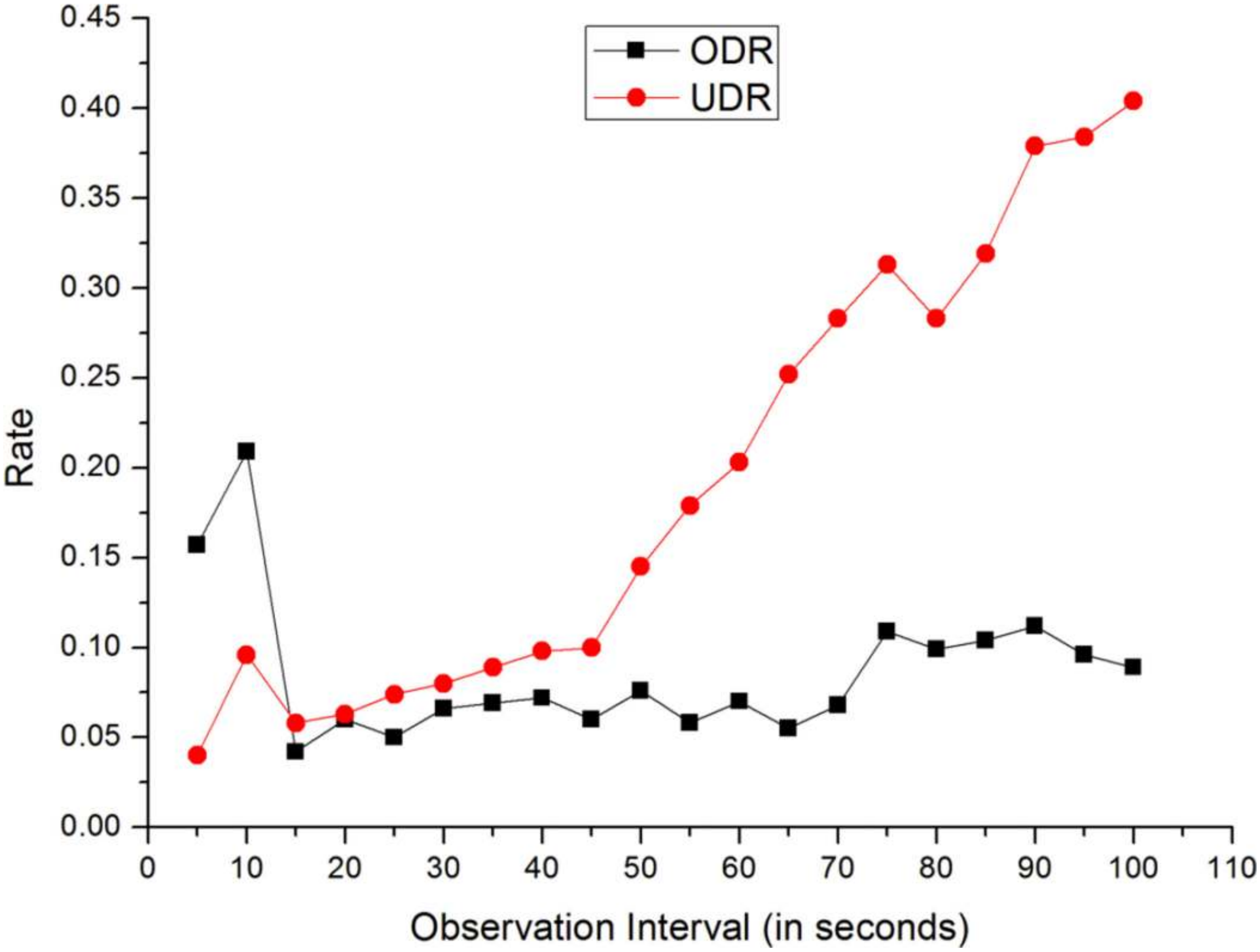

Figure 5.

Change of over detection rate (ODR) and under detection rate (UDR) with the increase of GPS observation interval

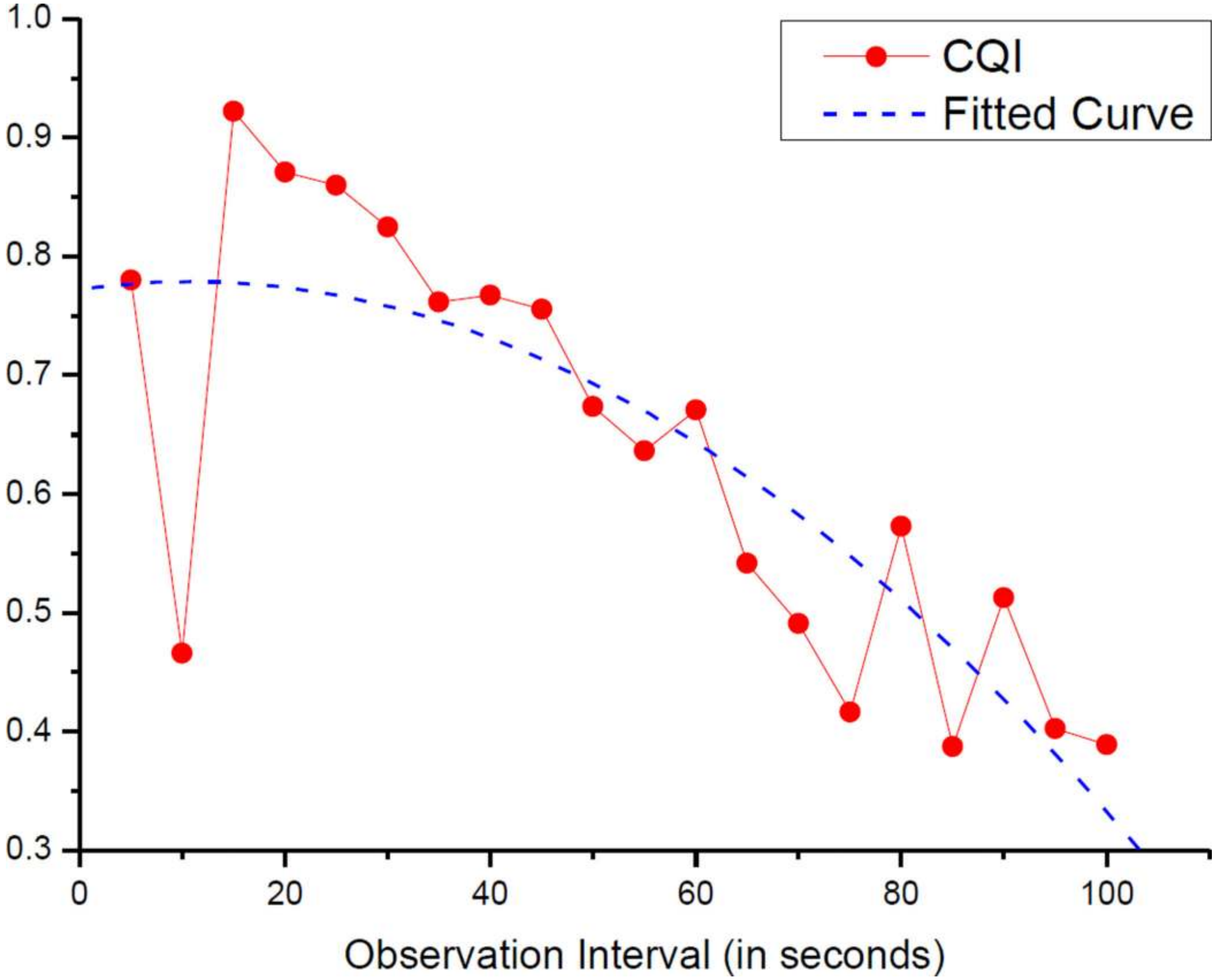

Figure 6.
Change of CQI with the GPS recording interval

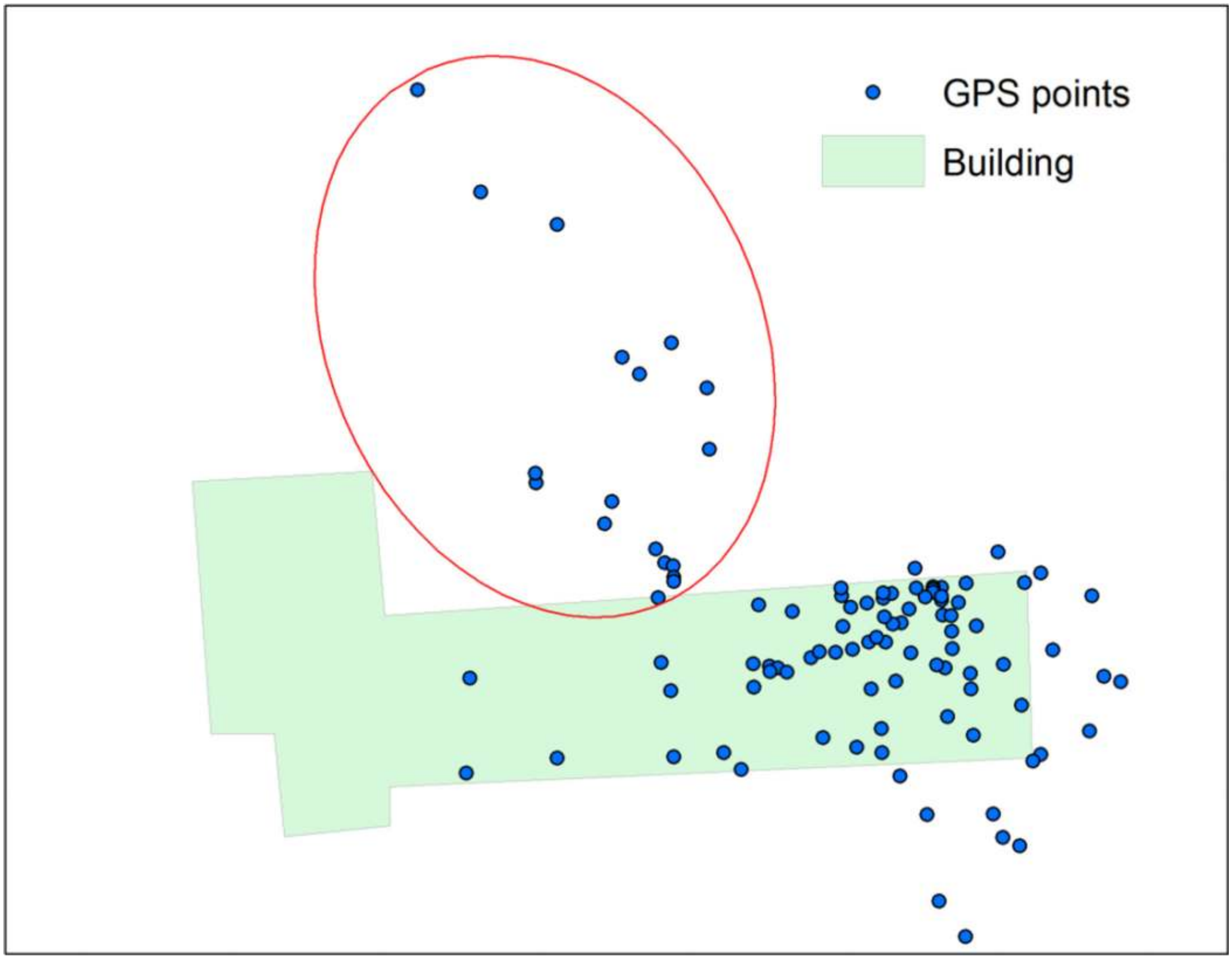

Figure 7.

An example of an over detection of walking activity: all GPS points in the figure belong to a staying activity inside the building. However, due to signal shielding, a group of consecutive points exhibit a curve pattern (shown inside the red circle) that was mistakenly classified as walking activity by the classification method proposed in this paper.

Table 1

Fuzzy rules for point membership state estimation

	Zero speed	**Slow speed**	**Fast speed**
Narrow angle	Staying: high Walking: low Other transportation: low	Staying: high Walking: low Other transportation: low	Staying: medium Walking: low Other transportation: medium
Wide angle	Staying: medium Walking: medium Other transportation: low	Staying: low Walking: high Other transportation: low	Staying: low Walking: low Other transportation: high

(Note: each cell shows the descriptive probability for the three outcomes according to input variables of speed and angle)

Table 2

Characteristics of the collected GPS data by phone model

Phone Model	Number of Collected GPS Points	Number of Observation Days	Number of Baseline Activities
Samsung Galaxy Note I	9,544	17	145
Motorola Droid X	29,992	15	192
Nokia N900	49,297	14	111

Table 3

Quality of detected activities of all phone models among different sampling intervals

Sampling Interval (seconds)	Number of identified activities	Number of reference activities	Over detection rate [a]	Under detection rate [a]	Mean location deviation (in meters)	Mean duration deviation (in minutes)	Classification quality index
5	638	448	0.157	0.040	0	3.366	0.7801
10	693	448	0.209	0.096	7.361	5.073	0.4659
15	474	448	0.042	0.058	10.64	3.307	0.9222
20	486	448	0.060	0.063	12.97	3.491	0.8711
25	463	448	0.050	0.074	15.1	3.583	0.86
30	470	448	0.066	0.080	15.69	3.665	0.8248
35	465	448	0.069	0.089	19.61	4.18	0.7615
40	469	448	0.072	0.098	18.02	4.004	0.7673
45	448	448	0.060	0.100	20.12	4.279	0.7554
50	419	448	0.076	0.145	20.19	4.714	0.6734
55	380	448	0.058	0.179	24.21	4.905	0.6363
60	359	448	0.070	0.203	22.28	4.042	0.6707
65	308	448	0.055	0.252	21.45	5.752	0.5419
70	295	448	0.068	0.283	22.2	5.924	0.4907
75	294	448	0.109	0.313	29.38	5.613	0.4165
80	353	448	0.099	0.283	23.38	3.926	0.5729
85	316	448	0.104	0.319	32.92	5.653	0.3873
90	267	448	0.112	0.379	25.85	3.501	0.5127
95	260	448	0.096	0.384	33.06	4.802	0.4024
100	246	448	0.089	0.404	28.73	5.192	0.3887

[a] Over detection rate: calculated by dividing the number of over detection cases by the number of identified activities;

[a] Under detection rate: calculated by dividing the number of under detection cases by the number of baseline activities.