Internet of Things 13 (2021) 100035

Contents lists available at ScienceDirect

Internet of Things

journal homepage: www.elsevier.com/locate/iot

Research article

Location prediction using GPS trackers: Can machine learning help locate the missing people with dementia?

Janusz Wojtusiak*, Reyhaneh Mogharab Nia

Machine Learning and Inference Laboratory, George Mason University, United States

ARTICLE INFO

Article history:
Received 1 October 2018
Accepted 6 January 2019
Available online 22 January 2019

ABSTRACT

Significant number of people with dementia are at risk of wandering and getting lost. These individuals may get hurt, cause distress to families and caregivers, and require costly search parties. This study explores the possibility of using machine learning methods applied to data from GPS trackers to create individualized models that describe patterns of movement. These patterns can be used to predict typical locations of individuals with dementia, and to detect movements that do not follow these patterns and may correspond to wandering. Data from a sample of 337 GPS trackers were used. After pre-processing the data are used for iterative clustering, followed by classification learning. The number of clusters ranged between one (devices that always stayed "home") and nine for devices with maximum mobility. The average number of clusters was 2.62. Models for predicting location achieved varying accuracy, depending on regularity of wearer's schedule. The achieved average Area under ROC (AUC) is 0.778, with accuracy 0.631, precision 0.662, and recall 0.604. Unusual locations that potentially correspond to wandering incidents were identified by applying a secondary classification learning after filtering out data corresponding to normal movement.

1. Introduction

Alzheimer's Disease (AD) and other forms of dementia constitute important public health concern. There are currently about 5.4 million individuals with dementia in the United States, with 70-80% of all people with dementia in the US being cared for at home by a family member [1] and 15 million caregivers provided annually an estimated 18.2 billion hours of care. In Virginia alone, there were approximately 447,000 family caregivers in 2013 and an estimated 455,000 in 2015 [2]. It is estimated that 60% of people with dementia will wander [1]. Wandering is a broad term that can be defined as "a syndrome of dementia-related locomotion behavior having a frequent, repetitive, temporally-disordered and/or spatially-disoriented nature that is manifested in lapping, random and/or pacing patterns, some of which are associated with eloping, eloping attempts or getting lost unless accompanied" [3]. The wandering may be a result of a person with a dementia type such as Alzheimer's not being able to remember his or her name or address, and becoming disoriented even in familiar places. *In the presented research, we focus on a specific aspect of wandering, that is being lost outside of home*. Wandering and getting lost can occur during the mild, moderate or severe stages of AD and is potentially dangerous (leading to falls and fractures, institutionalization and death) and may cause significant stress for families and caregivers [1,4]. Characteristics and behaviors associated with wandering include having dementia for a longer duration, severity of dementia (though wandering

* Corresponding author.
E-mail address: jwojtusi@gmu.edu (J. Wojtusiak).

https://doi.org/10.1016/j.iot.2019.01.002
2542-6605/

Table 1
Comparison of tracking devices available on the market.

Device	Technology	Indoor/Outdoor	Style
TRX Systems	Beacons, Ultra-wideband	Indoor	Wearable
Yepzon One	Bluetooth, GPS, GSM	Indoor, Outdoor	Necklace
Amcrest	GNSS/GPS, GPRS/GSM	Outdoor	Portable
Trax	GPS, GLONASS	Outdoor	Carriable
Americaloc GL300W	GPS, GSM	Outdoor	Carriable
AngelSense	GPS, GSM	Outdoor	Attachable
GPS SmartSole	GPS, GSM	Outdoor	Shoe sole
Mindme	GPS, GSM	Outdoor	Pendant
Safe Link	GPS, GSM	Outdoor	Carriable
Spy Tec	GPS, GSM	Outdoor	Attachable
Project Lifesaver	RF	Outdoor	Bracelet
Q-Track NFER RTLS	RF, Bluetooth, WLAN	Indoor	Attachable
Accuware	Wi-Fi, Beacons	Indoor	Carriable/Attachable
infsoft	Wi-Fi, Beacons, Ultra-wideband, RFID	Indoor	Carriable/Attachable
iTraq	Wi-Fi, GPS, GSM	Indoor, Outdoor	Attachable
MX-LOCare	Wi-Fi, GPS, GSM	Outdoor	Watch
PocketFinder	Wi-Fi, GPS, GSM	Indoor, Outdoor	Attachable
Mini A9 GPS Tracker	Wi-Fi, LBS, GPS, GPRS/GSM	Indoor, Outdoor	Necklace

can occur at any stage), presence of a sleep disorder, impairment in day-to-day functioning, and behavioral disturbances such as anxiety and depression [5]. Recent research on the management of wandering behavior focuses on promoting safe walking which often includes electronic tagging of a person who wanders. While GPS tracking of people with AD is often seen as unethical because it decreases a person's autonomy and the individual's right to privacy, there are no alternatives except of constant supervision [5,6].

Ubiquitous presence of health trackers, GPS devices, smartwatches and other wearable IoT technologies open new possibilities for improving safety and care for individuals with AD. The purpose, function and technology offered by these devices vary, and ranges from gait and movement analysis to assess physical activity, to alert systems that detect falls, to GPS trackers that help locate the missing. There are multiple trackers available on the market now, some of which are advertised specifically for the elderly or individuals with AD, including MX-LOCare, Mindme, Pocketfinder, GPS Shoe and its successor GPS SmartSole, to mention just few. The latter technology is used in this research as it provides real time monitoring of wearers. There are also other technologies such as one used by Project Lifesaver that uses radio location to help find the missing individuals (projectlifesaver.org). Table 1 compares features of some of the popular devices available on the market.

There are a number of previous research project related to the approach presented here. The closest available research is by Shoval et al. [7,8] in which the authors identified differences in movement patterns between people with mild dementia, MCI (Mild Cognitive Impairment) and no cognitive impairment. They found out that participants who suffer from mild dementia have much less varied mobility patterns than healthy participants and those with MCI and they usually stay close to their homes. People with dementia go out at routine times, although it varies from person to person: some stay in the familiar surroundings while others move farther. However, in their approach, the authors did not consider prediction and detection of wandering patterns. According to a video-based observational study conducted by Martino-Saltzman et al. [9], patients follow four basic travel patterns: direct trajectory, pacing, random trajectory and lapping. The latter 3 patterns are categorized as wandering. Vuong et al. [10,11] used this rule to implement a classification algorithm for detecting these trajectories. In another experiment, the same team defined a feature vector for each trajectory including displacement, path length, total travel time, average velocity, straightness index, directional mean, and circular variance, and used existing supervised learning algorithms to classify trajectories into the 4 categories. They achieved the best accuracy of 72% using a Random Forest classifier. In another related research, Delaunay and Guérin [12] detected wandering behaviors from movement patterns using GPS data collected by GPS watch trackers. Their algorithm uses the metrics coming from the GPS and sends an alert only when all the wandering behavior patterns are detected. Lin et al. [13] proposed a real-time method to detect wandering behavior by finding adjacent turning points (points in a trace with a vector angle equal to or more than 90 degrees) within a distance range which form loop-like movements. Sposaro et al. [14] built an Android application to detect wandering patterns which uses two-phase approach proposed by Yin et al. [15] in which a one-class Support Vector Machine (SVM) model filters out normal data, then the abnormal activities are detected from a normal activity model using kernel nonlinear regression (KNLR). The data collected was then evaluated by Bayesian networks which determine the probability of wandering behavior. Kearns et al. [16] found a link between dementia diagnosis and path tortuosity (change of movement path) recorded in an indoor assisted living facility using Fractal Dimension approach. Tung et al. [17] used GPS data to measure life space of individuals with AD. There are many other examples of applications of machine learning methods to GPS data and prediction of location is relatively well established (i.e., [18–23]).

The approach presented here has many similarities to research previously done by others, but is distinct in many ways. Our main objective was to create models capable of predicting person's location, both during routine schedule and when lost. This is achieved by a multi-stage process that involves unsupervised and supervised learning.

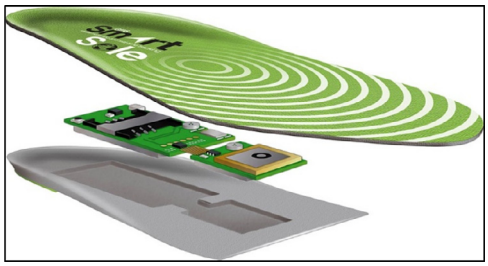

Fig. 1. SmartSole design (source: gpssmartsole.com).

The remainder of this paper is organized into two main components, description of data and methods, analysis of obtained results, are followed by conclusion and research directions in GPS tracker data analysis.

2. Method

In the presented work, machine learning (ML) techniques are used to identify frequent wandering and detect spatiotemporal patterns that may allow for prediction of future incidents. The overall idea behind the method is that one needs to start with predicting typical locations, and then move to identifying and predicting anomalous ones that may correspond to the wandering incidents in which patients get lost. This approach directly corresponds to what is used by search & rescue community.

1. Identify most likely general area (or ranked list of areas) where the person is likely to be, from the list of frequent locations.
2. If the person is not at the location (i.e., home, daycare facility), employ secondary models that predict potential trajectories of movement of person from that location(s) based on known movement patterns, including previous wandering incidents.

The specific work presented here focuses on the first step of identifying and predicting frequent locations. To do so, two types of ML methods are used. Unsupervised learning algorithms are used to find patterns in the data that correspond to frequent locations in which GPS tracking devices are typically located. Supervised learning methods are used to construct models for predicting likely locations.

There is an important limitation of the proposed work. The data analyzed here did not include any information about the device users and their dementia status, therefore we cannot claim any relationship between obtained results and confirmed incidents of wandering. Our assumption is that the wearers are the elderly with dementia, which is the main market for the GPS trackers used, but this is not guaranteed. We also had no control over how often the devices were worn, charged, and who managed them.

2.1. GPS SmartSole

SmartSoles by GTX Corp. are tracking devices that include GPS and GSM units embedded in shoe insoles to provide real-time geolocation data for wearers. The company also offers monitoring service, and notification of events in which wearers cross pre-defined geozones. The SmartSoles (Fig. 1) fit many types of shoes and can be adjusted to wearers providing additional comfort (gpssmartsole.com). Because these devices are “hidden” inside shoes, they reduce chances of being discarded by individuals suffering from dementia known to remove foreign objects. The data used in this project has been obtained from GTX Corp. as part of data use agreement between the company and our research group. It is important to note that our research group has no interest in the GTX Corp. other than access to data for research that is not funded by the company in any way.

2.2. Spatiotemporal GPS Data

The data consist of flat data tables that include the following fields: device ID, GPS timestamp, server timestamp, longitude, latitude, device status. On the first look the data for a single GPS device are very simple as only longitude, latitude, and time stamp matters. In practice, GPS data are very complex once spatiotemporal relationships between points are considered. The GPS data are potentially high frequency irregularly spaced spatiotemporal, with many missing observations. In the presented work we applied a simple transformation to change spatiotemporal data into weighted spatial data that correspond to devices being stationary over a given period of time (see method description below). That transformation allowed for application of clustering and classification.

From the obtained dataset, a sample of 337 devices with at least 14 days of recorded data was selected. Location and movement patterns were analyzed separately for each device, independent of other devices, thus creating individualized

Fig. 2. Device data timeline.

Table 2
Device status statistics.

Sample size: 337	Mean	SD
Device status		
% of time in motion	9%	10%
% of time not moving	84%	18%
Other device status		
% of time arrived at geozone	2%	8%
% of time departed geozone	1%	2%
% of time battery is low	5%	12%
% of time spent at home	52%	31%
Number of data points	5572.47	7390.75
Number of days	99.54	125.39
	Grand Mean	**Grand SD**
Mean distance of device from home	29.37 km	195.32 km

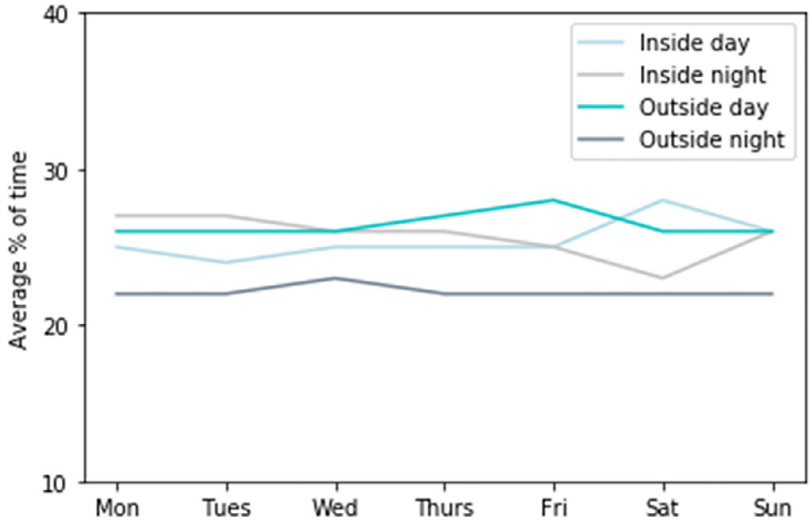

Fig. 3. Average percentage of recorded time spent in and outside during the daytime and nighttime.

models. In the experiments, at least last 7 days of data were used to test models and the reminder of data for training. The rationale for using time-based split is that models are intended to predict user behaviors in the future, and this is not guaranteed with random split between training and testing. This is illustrated in Fig. 2. Brief summary of the data is shown in Table 2 and Fig. 3.

Table 2 shows distribution of status codes for the devices and Fig. 3 shows the average percentage of recorded time spent in and outside during daytime and nighttime. From the data, one can immediately see that there is no significant difference between use of devices in different days of week. Few devices have setup geozones that are used to alert caregivers when the device leaves area, thus time spent at geozones is significantly smaller than time spent at home location.

2.3. Data analysis

The approach described below has been created to handle spatiotemporal nature of the GPS data, detect patterns of typical movement and usual locations. These usual locations are baseline for further detection of unusual locations/movements that can be consequently linked to patient wandering. The complete procedure is outlined below. To create patterns of movement for individuals with AD, raw GPS data need to be normalized and transformed before machine learning algorithms can be applied.

Frequent locations are discovered by applying clustering algorithms, followed by classification learning needed to predict frequent locations, and labeling of data not belonging to typical locations as suspected wandering incidents. Finally, the unusual locations can be confirmed by secondary classification.

Step 1: Select devices and data

In order to analyze patterns of movement for devices with sufficient activity, the process starts with removing data for devices that: (a) were never activated outside of manufacturer's facility; (b) have less than 14 days of data. The minimum of 14 days of data is required later in the prediction step, 7 days for training and 7 days for testing. This is followed by removal of erogenous data points: (c) data that correspond to the manufacturer location. These are typically first few minutes of the device's recorded activity and in some cases towards the end. For SmartSole devices these are at the company location in Los Angeles, CA. Finally, (d) remove erroneous data, such as (0,0) coordinates, and those with dates in the past. The resulting dataset is referred to as D. For simplicity we assume that $D = \{(lat_i,\ lon_i,\ t_i)\ i = 1,2,\ldots\}$ that correspond to time-stamped latitude and longitude of GPS locations.

Step 2: Determine home location

A simple approach was used to calculate home location for each device as the most frequent 100×100 feet rectangle present in the data D. The size of the rectangle was chosen because of typical GPS accuracy and household size. There are other approaches to the problem of discovering home location known in the literature, but these methods do not provide better detection for the specific dataset we used. For example, Lin and Cromley [28] provided comparison of three methods for discovering frequent locations based on Twitter data. The authors concluded that weighted most frequently visited approach is the most accurate.

The advantage of the simple approach used in the presented work is that it can be easily done directly within database containing data and do not require use of any additional software.

Step 3 (optional): Deidentify data

GPS data are highly sensitive and need to be protected [6,27]. In order to use data for research purposes, one can follow a simple deidentification procedure. The process is to normalize and transform data D –> D_{Norm}, so that the home location is in (0,0) coordinates, and other locations are randomly rotated by the same angle. The angle of rotation is randomly selected for each device. Such transformation makes it computationally very expensive to discover real location, but preserves relationships between points thus allowing for further analysis.

Step 4: Convert temporal data to weighted frequency domain

GPS data are irregularly spaced in time. In order to perform location analysis one needs to estimate how long a person stays in a given location. The approach used here is to convert timestamped geolocation data to *weighted frequency domain* in which weight of each data point is assigned as

$$w_f(x_i) = \begin{cases} \frac{t_{i+1} - t_i}{c} & if\ dist(x_i, x_{i+1}) <= \lambda\ \wedge\ \neg s(x_{i+1}) \\ 1 & otherwise \end{cases}$$

where t_i and t_{i+1} is time measured at points x_i, x_{i+1}. In the data, by default the time is measured in seconds. We considered a constant c to be 600 s (10 min) that correspond to typical data frequency in GPS trackers used. A data point is weighted if its distance to the next point is less than λ. Here, we used $\lambda = 0.05$ km after conducting experiments with different possible values (see parameter tuning section). This approach weights only data when the device is stationary. Additionally, to prevent counting time when the device is turned off we considered $s(x_{i+1})$ that indicates status of the device is switched off. The raw data do not include on/off status of the devices; therefore, a device is considered off if $t_{i+1} - t_i > 3900$ s. The value has been determined experimentally based on distribution of the data discussed later in parameter tuning section.

Step 5: Split data into training and testing sets

The approach previously shown in Fig. 2 is used to split data into training and testing. In the presented experiments, data available at the end of each device's recorded period is used for testing. The split into training and testing in this fashion is reasonable as it reflects real use of the devices for which models are being trained for a certain period of time, and then models are applied. In fact, this approach is more reasonable than, for example, 10-fold cross-validation that does not preserve ordering of data when splitting into training and testing.

In a real-world application of the described method we envision a system that will progressively learn models as more data are available, with majority of data used for training and last period used for evaluation of models created "so far". When quality of the models achieve a desired level the models are considered to be useful for prediction. It is also possible that for a long-time wearers of the devices old patterns/data need to be forgotten to encounter for movement pattern drift.

Step 6: Detect frequently visited locations

There is a large number of publications that describe different approaches to finding frequent locations based on geolocation data, including GPS. Ashbrook and Starner [32] ran a variation of k-means clustering algorithm multiple times to find an optimal radius by looking for a knee on a graph plotted based on the results, i.e. a point of time where number of places

Algorithm 1 Pseudocode of DBSCAN-MP, multi-pass clustering method.

Input: D, α_i, β_i, ε_i
Output: $C = \{C_1, \ldots C_k\}$
1. $i := 1$; $D_1 = D$
2. **REPEAT**
3. $MinPts = \begin{cases} \alpha_i * \sum_i w(x_i) \text{ where } x_i \in D_i \text{ if } \alpha_i * \sum_i w(x_i) > \beta \\ \beta \text{ otherwise} \end{cases}$
4. $C_i := DBSCAN(MinPts, \varepsilon_i, Di)$
5. $D_{i+1} = D_i \setminus C_i$
6. $i := i + 1$
7. **UNTIL** $C_i = \emptyset$
8. Output $C = \bigcup C_i$

start to shoot up. There are however drawbacks with this method. The number of clusters should be known before clustering starts. Also, the final clustering is dependent on the initial random assignment of points to clusters. Not surprisingly, according to the literature, most approaches use clustering techniques based on the point density which overcome shortcomings of k-mean clustering methods. Isaacman et al. [29] identified significant locations using Hartigan's leader method which is a spatially clustering algorithm [30] and applying logistic regression model on CDR (Call Detail Record) data. By far the most popular algorithm applied to GPS data is Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and its variants [25]. Tuo et al. [31] designed SSMA (Speed and State based Mining Algorithm) approach in which DBSCAN clustering and naïve Bayes classifier methods are applied on cellular data to discover important places. Zhou et al. [38] developed DJ-Cluster clustering algorithm which is density-based but simpler than DBSCAN despite sharing similar definitions to overcome DBSCAN performance problems. However, to avoid the complexity problem in DBSCAN, R-Tree based index can be used which would result in complexity same as DJ-Cluster [33].

In the described research we used DBSCAN-MP (Multi-Pass), an extension to the popular DBSCAN algorithm that is particularly applicable to geospatial data. DBSCAN groups adjacent points in dense areas, that is points whose distance is less than or equal to specified distance into a cluster as long as group size reaches the minimum specified cluster size. Otherwise, points are marked as outliers (low-density regions). The used approach DBSCAN-MP is specifically designed to handle geolocation data that is highly imbalanced (i.e., there is a very large difference between cluster sizes). When clustering such data, it is difficult to properly set parameters of DBSCAN so that both large and small clusters are properly discovered. The algorithm starts with discovering a small number of large clusters, removes them from the data, and interactively continues to discovery of smaller and smaller clusters. The process is shown in the Algorithm 1 below:

The presented algorithm that applies clustering multiple times is needed because of large disproportion of weights between the most common (home) location, followed by other large clusters, and other locations, which also varies between devices. Without the iterative application of clustering, it is practically impossible to select *MinPts* (parameter that defines minimum cluster size) that result in reasonable clustering. The above method relies on DBSCAN particularly applicable to geospatial data, but any other clustering algorithm can be used instead. We have experimented with other clustering methods, including OPTICS [26], k-means, and hierarchical clustering, but decided on using BDSCAN because of its superior accuracy. Another advantage of DBSCAN is that it does not force every point to belong to a cluster, but instead can label points as noise in areas that are not dense. This property is specifically important in the following step.

Step 7: Construct models for predicting typical locations

Prediction of typical location is a natural application of classification models with output representing cluster in which person is likely to be. In order to do so, one needs to construct models for predicting typical locations belonging to clusters C_1, … C_K given known information about a person with AD. The simplest form of known information is given by day of a week and time. That can be supplemented by a set of previously visited locations if known, distribution of time spent at different locations within a given time window, and other derived attributes.

The first step in the classification learning is to remove the points between clusters. These points are marked as noise in the clustering step. Also it is important to remove the data related to the most frequent location (home) since it comprises most of the points and would make the prediction biased. The resulting models can be used for location prediction assuming the lost person with AD is not at home.

Such prepared data are ready for location prediction in which ML is applied to create classification models for predicting normal location (given as a cluster/normal location $C_1..C_K$) given day and time, along with additional derived attributes. In our initial experiments, we have applied several classification learning methods from which Random Forest [24] shows the best results.

Step 8: Label unusual trajectories

This step is to label the data that do not belong to clusters or routes between clusters. We make a simple assumption that such data represent routes/trajectories in which both origin and destination are in the same cluster. Because most of wandering incidents happen at home, most of such trajectories are around home location. These trajectories can be marked

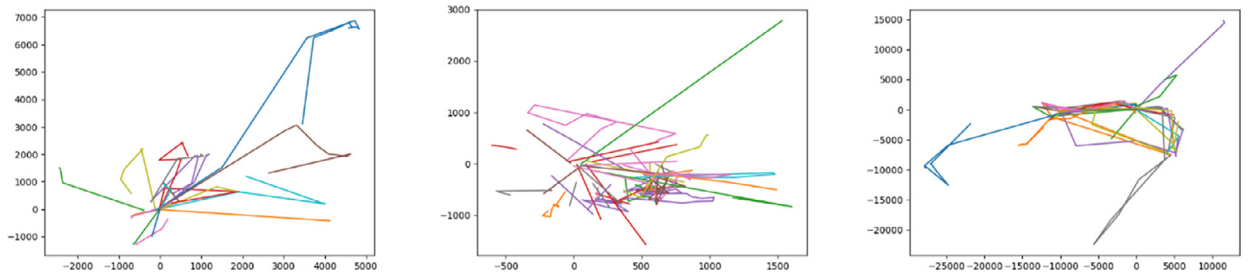

Fig. 4. Movement trajectories around home location for three example devices. All distances are shown in meters.

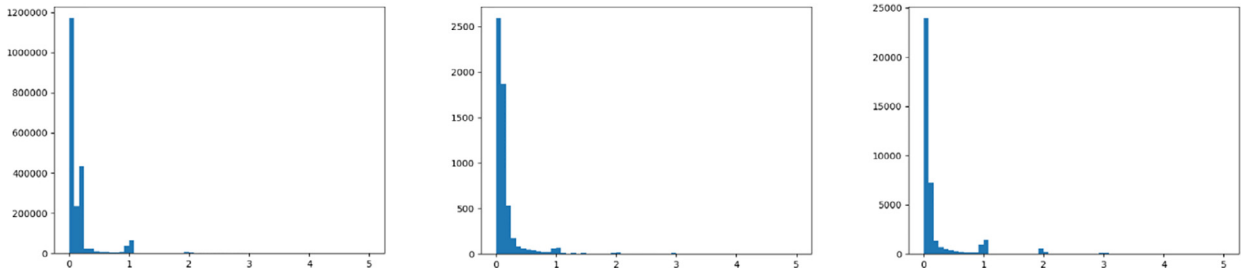

Fig. 5. Example frequencies of time gaps less than 5 h for all the devices (left) one randomly selected device with approximately 5000 data points (middle); and one device with approximately 40,000 data points (right).

as potential wandering incidents m_1, m_2, ... m_p. If additional data are available that report on confirmed incidents of getting lost, these incidents can be further narrowed to distinguish between getting lost and normal movements around a single cluster. Fig. 4 illustrates example routes around home location for three example devices.

Step 9: Model unusual locations

Once trajectories around home are identified, it is possible to create models for predicting location within a given trajectory. This can be viewed as movement analysis on a local scale, as compared to global analysis when predicting usual locations. There is an extensive literature available on clustering and prediction of locations from trajectory data. For example some recent work include those of Bian et al. [34] that presented a review of trajectory clustering methods that they generally categorized into unsupervised, supervised, and semi-supervised. Wang et al. [35] as well as Cai et al. [36], analyzed vehicle trajectory data to discover anomalies in taxi movements. Tasnim et al. [37] incorporated semantics into the process of sub-trajectory detection. There are many other related works.

The above Steps 8 and 9, were presented for completeness, but are out of scope of main experimental evaluation reported in this paper. Here we focus on detection and prediction of usual locations.

2.4. Parameter tuning

In order to choose the best possible parameters needed for our algorithms, we executed them with different combination of parameters needed in application of DBSCAN-MP, as well as for classification learning of typical locations.

To obtain threshold value for time gaps up to which the device could be assumed continuously on, indicted as $s(x_{i+1})$ in Step 4 above, we analyzed frequencies of temporal distance between consecutive points reported by GPS trackers. These data are illustrated in histograms (Fig. 5) that shows averages for all devices as well as two randomly selected ones. The analysis indicated that 3900 s is an optimal value to be used as a threshold. Note that the work can be extended by adaptively selecting a device specific threshold that would account for difference in usage patterns between different users.

In order to select the optimal parameters used by the clustering algorithm, we performed optimization that searched over large number of combinations of possible values as follows:

- $\boldsymbol{\lambda}$: maximum distance between two data points in order to be considered indistinguishable. The parameter is used in conversion from temporal to weighted frequency domains. The tested values are $\boldsymbol{\lambda} \in \{0.05, 0.1, 0.2, 0.3, 0.5, 0.7, 0.9\}$.
- $\boldsymbol{\varepsilon}$: maximum distance between two points in order to be considered as neighbors in the DBSCAN clustering algorithm with possible values of $\boldsymbol{\varepsilon} \in \{0.05, 0.1, 0.2, 0.3, 0.5, 0.7, 0.9\}$.
- $\boldsymbol{\alpha_1}$: minimum support for the first DBSCAN with potential values of $\boldsymbol{\alpha_1} \in \{0.1, 0.15, 0.2, 0.25, 0.3\}$.
- $\boldsymbol{\alpha_2}$: minimum support for the second DBSCAN with possible values of $\boldsymbol{\alpha_2} \in \{0.01, 0.02, 0.03, 0.05, 0.08, 0.1\}$.

First, we experimented with values of $\boldsymbol{\varepsilon} \in \{0.3, 0.5, 0.7, 0.9\}$, $\boldsymbol{\alpha_1} \in \{0.1, 0.15, 0.2, 0.25\}$ and $\boldsymbol{\alpha_2} \in \{0.01, 0.05, 0.08, 0.1\}$. Smaller values of $\boldsymbol{\varepsilon}$ and $\boldsymbol{\alpha_2}$ resulted in higher accuracy whereas in the first DBSCAN iteration, minimum sample support ($\boldsymbol{\alpha_1}$) with the bigger values had better results. Therefore, we chose $\boldsymbol{\varepsilon} \in \{0.05, 0.1, 0.2\}$, $\boldsymbol{\alpha_1} \in \{0.2, 0.25, 0.3\}$ and

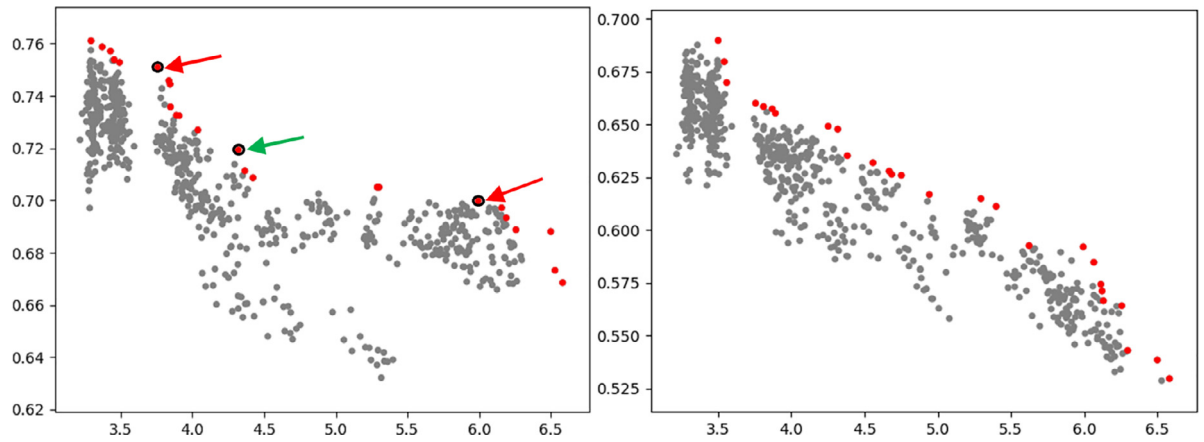

Fig. 6. Average number of locations vs. average AUC (left) and accuracy (right). Pareto solutions are marked in red. The solutions indicated with arrows are selected for further experimentation.

Table 3
Cluster statistics based on 337 devices. Note that the statistics include devices for which the method was not able to identify any additional clusters besides home.

	Mean	SD
Number of clusters	2.62	1.73
Number of clusters excluding home	1.62	1.73
	Grand Mean	Grand SD
Mean number of days spent in clusters	41.78	51.20
Mean number of days spent in clusters excluding home	5.74	13.78
Mean number of data points in clusters	2242.91	3494.02
Mean number of data points in clusters excluding home	247.24	424.02

$\boldsymbol{\alpha_2} \in \{0.01, 0.02, 0.03\}$ for the second set of experiments. After running the clustering step with the above values and building our predictive model in the next step, we ran the pareto frontier optimization algorithm and arrived at a number of satisfactory solutions that provide a compromise between AUC and number of clusters (that allows for more precise location detection). The pareto optimal solutions chosen can be seen in Fig. 6.

The results shown above indicate three potentially useful combinations of parameters for further investigation. The final set of parameters indicated with green arrow in Fig. 6 that correspond to values of $(\boldsymbol{\lambda}, \boldsymbol{\varepsilon}, \boldsymbol{\alpha_1}, \boldsymbol{\alpha_2}) = (0.05, 0.2, 0.3, 0.03)$ is selected.

In order to select best method for classification learning of usual locations, logistic regression, naïve Bayes, Random Forest, and Support Vector Machine algorithms were tested. Random Forest has been executed with number of trees ranging from 50 to 1000. After analysis of results, RF with 100 trees has been selected because larger number of trees do not improve results.

3. Results

The method described in the previous section has been applied to a sample data obtained from GTX Corp. In this section, the obtained results are reported as follows. First, results of clustering are presented, followed by prediction of usual locations. Then, early work on results of classification of unusual locations is shown. Finally, selected results that indicate possibility of detecting change of patterns over time are described. The results are based on data being split into training and testing as outlined earlier in the paper: the last 7 days or 25% of data, whichever includes more days, was set aside for testing, and remainder of the data used for training.

3.1. Clustering

A typical data of location with marked (in color) frequently visited regions is presented in the Fig. 7 below. Note that the location coordinates are recalculated so that home location is at (0,0) coordinates. The plots on the left represent clusters obtained in step 5 of the above method and plots on the right represent complete set of clusters. Table 3 shows overall statistics of the clusters constructed with parameters of $\lambda = 0.05$, $\varepsilon = 0.2$, $\alpha_1 = 0.3$ and $\alpha_2 = 0.03$ (one of selected Pareto solutions).

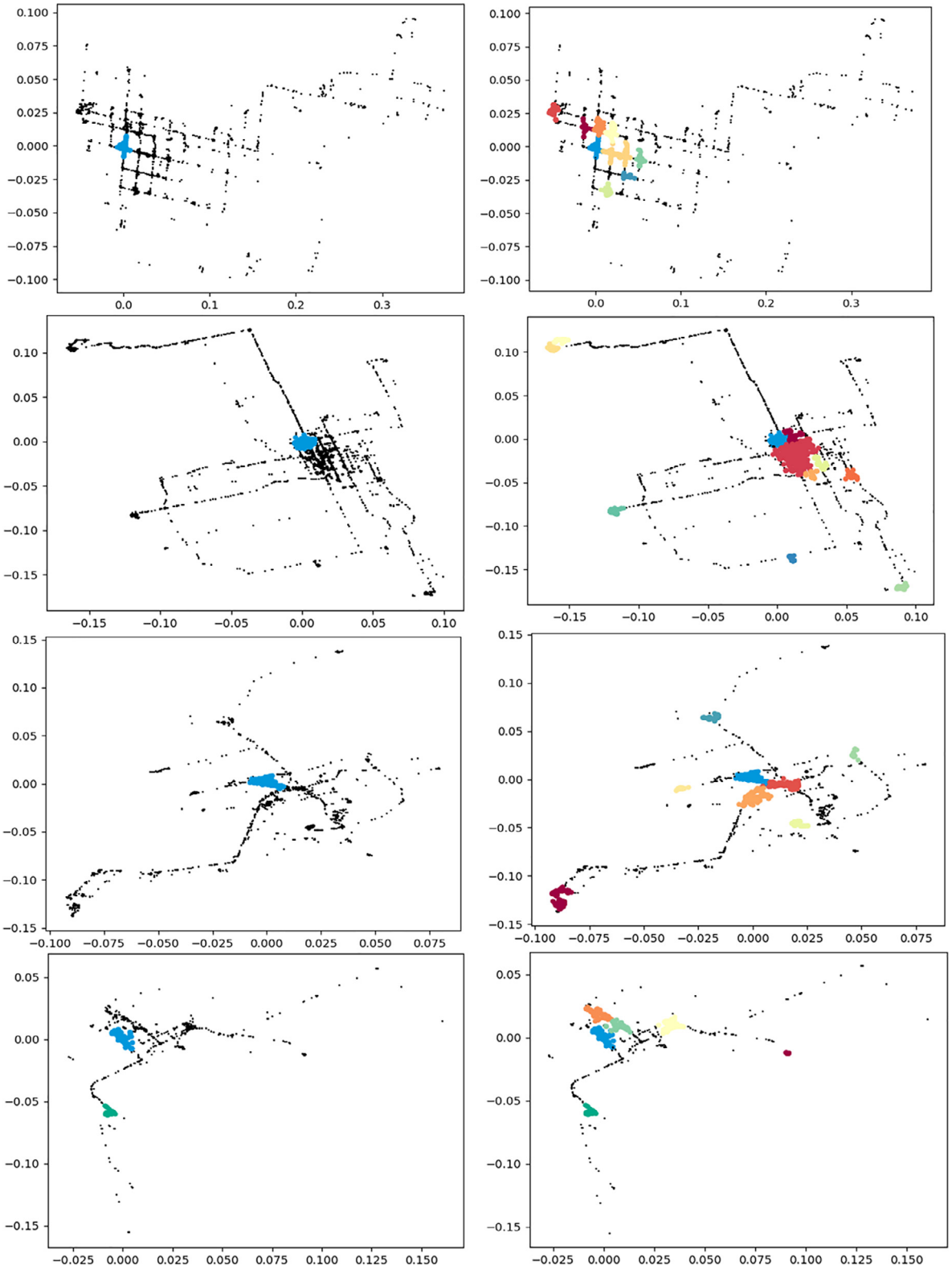

Fig. 7. Identified frequently visited locations for six GPS trackers. Left: clusters created in DBSCAN-MP pass 1; Right: complete set of clusters from DBSCAN-MP pass 2.

Table 4
Prediction of usual movements statistics.

AUC (% of devices)	AUC <= 0.60 (29%)		0.60 < AUC <= 0.75 (26%)		AUC > 0.75 (45%)		Total	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Accuracy	0.45	0.27	0.48	0.20	0.77	0.20	0.60	0.28
Number of clusters	4.06	1.43	4.73	1.31	4.54	1.65	4.32	1.53
Number of days	140.48	110.85	207.9	267.49	112.56	122.91	141.83	165.18
Number of data points	6096.91	7626.55	9320.47	10,450.88	8039.79	8221.51	7587.11	8342.76

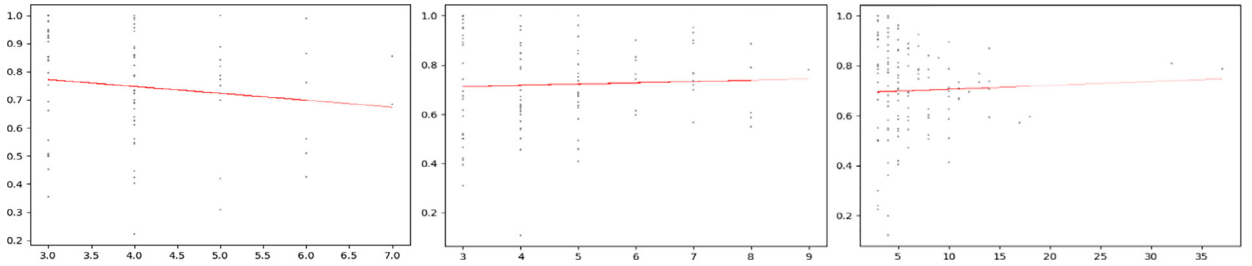

Fig. 8. AUC vs. # of frequent locations of devices for the chosen pareto-optimal solutions.
$\lambda = 0.1,\ \varepsilon = 0.3,\ \alpha_1 = 0.25,\ \alpha_2 = 0.05$; mean AUC = 75% (left)
$\lambda = 0.05,\ \varepsilon = 0.2,\ \alpha_1 = 0.3,\ \alpha_2 = 0.03$; mean AUC = 72% (middle)
$\lambda = 0.9,\ \varepsilon = 0.3,\ \alpha_1 = 0.1,\ \alpha_2 = 0.01$; mean AUC = 70% (right).

Fig. 7 illustrates partial and complete sets of clusters obtained after the first and second pass of DBSCAN-MP, respectively. As one can see, the second pass of the method identifies additional smaller clusters after deletion of large/dense areas from the data.

3.2. Prediction of usual movement

Next step after clustering is to apply supervised learning to construct models for predicting typical locations. According to initial results, the performance of the method depends on specific individual being followed.

In order to establish prediction baseline we built simple classification models based on only day of the week and time of the day as input attributes, and cluster assignment as output. Surprisingly, this simple approach is able to correctly predict 100% of outside home locations for certain devices, indicating a very regular lifestyle of wearers. For others, the accuracy can be as low as 0%. The three Pareto-chosen parameter sets (Fig. 6) lead to average AUC of 75%, 72% and 70%. Further investigation into devices with AUC of 0, indicates limited availability of data. This is illustrated in Fig. 11 for one specific device for which the clusters of frequent locations completely change month to month. It is an indication of a drastic pattern drift in the data beyond possibility of simple incremental update of models.

Table 4 reports average accuracies and statistics for prediction of locations using only day of the week and time of the day. The results indicate that higher accuracy (AUC) is typically related to larger number of available data points, but is unrelated to number of clusters of usual locations discovered as also shown in Fig. 8.

Further set of experiments was designed to test if knowledge of history of movement used as input to classification learning improves prediction. Two approaches were considered. First, an n-gram of past known n *locations* was inserted into data resulting in data in the form: (DW, T, C^1, C^2, ... C^n, C^{n+1}), where DW is day of week, T is time of the day, C^1, C^2, ... C^n are past known locations, and C^{n+1} is the (unknown) location being predicted.

Similarly, an experiment was performed that includes duration/proportion of time in top n clusters. The data were created in the form: (DW, T, D^1, D^2, ... D^n, C^{n+1}), where DW is day of week, T is time of the day, D^1, D^2, ... D^n indicate proportion of time spent in locations 1..n in the past, and C^{n+1} is the (unknown) location being predicted.

Finally, all of the above methods were applied to simulate a person being lost for a certain period of time up to 5 h. Results of the prediction are shown in the Tables 5 and 6 below, and visually depicted in Fig. 9. Delay 0 indicates prediction without any added delay to the data.

Interestingly, in several of the models, the additional information worsens results. Mean AUC becomes lower than one from simple use of day of the week and time of the day. This may indicate overfitting of models due to lack of sufficient data, but further analysis is needed to confirm that hypothesis.

3.3. Change of movement patterns

In order to check for reasons of inability to predict usual locations for some devices and very high accuracy for others, a simple comparison of patterns over time was done. Not surprisingly, comparison of clusters over time indicates that

Table 5
Results of prediction of usual location using different coding of inputs.

	AUC	Accuracy	Precision	Recall
DT	0.719	0.605	0.642	0.565
Location n-grams				
DT + 1-gram	0.763	0.627	0.670	0.597
DT + 2-gram	0.754	0.643	0.673	0.614
DT + 3-gram	0.762	0.641	0.672	0.610
DT + 4-gram	0.763	0.637	0.672	0.607
DT + 5-gram	0.757	0.635	0.660	0.600
Duration n-grams				
DT + 1-gram	0.775	0.638	0.661	0.612
DT + 2-gram	0.778	0.631	0.662	0.604
DT + 3-gram	0.774	0.634	0.669	0.608
DT + 4-gram	0.774	0.632	0.672	0.608
DT + 5-gram	0.767	0.626	0.659	0.596
DT + 6-gram	0.756	0.629	0.666	0.596

Table 6
Results of prediction of usual location using different coding of inputs and delay up to 5 h. The numbers indicate AUC of the models.

	0	1/2	1	3	5
DT	0.719	0.719	0.719	0.719	0.719
Location n-grams					
DT + 1-gram	0.763	0.733	0.720	0.719	0.724
DT + 2-gram	0.754	0.737	0.727	0.729	0.728
DT + 3-gram	0.762	0.735	0.727	0.724	0.724
DT + 4-gram	0.763	0.733	0.726	0.722	0.722
DT + 5-gram	0.757	0.728	0.727	0.719	0.723
Duration n-grams					
DT + 1-gram	0.775	0.728	0.716	0.716	0.722
DT + 2-gram	0.778	0.741	0.730	0.717	0.722
DT + 3-gram	0.774	0.727	0.718	0.715	0.724
DT + 4-gram	0.774	0.734	0.715	0.706	0.713
DT + 5-gram	0.767	0.734	0.710	0.701	0.713
DT + 6-gram	0.756	0.727	0.706	0.696	0.708

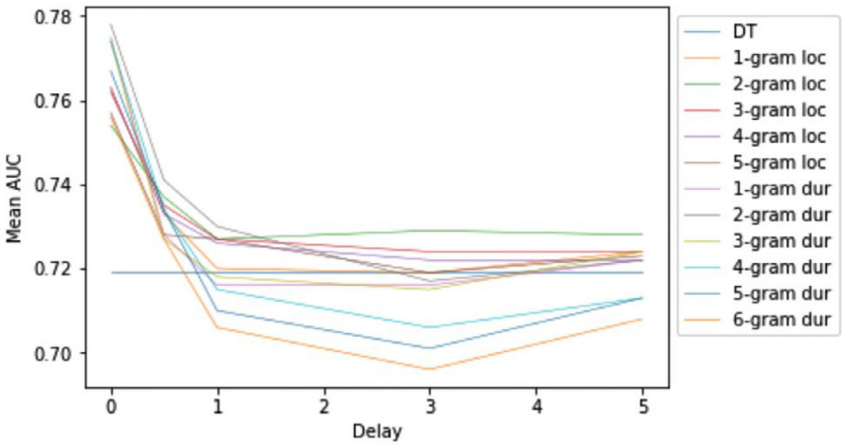

Fig. 9. AUC of models for predicting usual locations.

movement patterns are stable for the majority of devices, but drastically change for others. Hotelling's_T-Squared test was used to compare data in clusters over time. The obtained p-values were high (>0.1) which indicates that clusters representing movement patterns are consistent over time (Fig. 10). However, as indicated earlier and depicted in Fig. 11, a small portion of devices record data with radically different pattern of movement over time. Further work and additional data are needed to evaluate reasons for that irregularities.

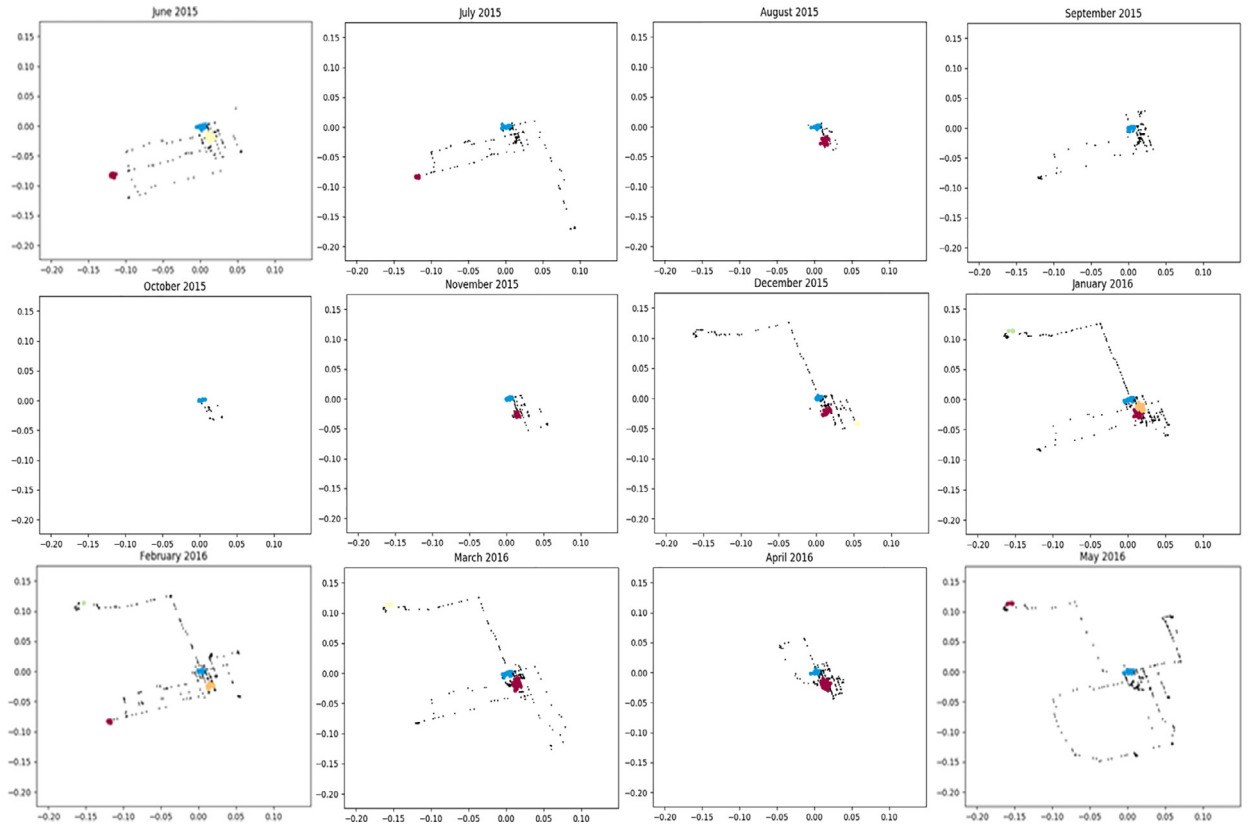

Fig. 10. Clusters over time for an example device.

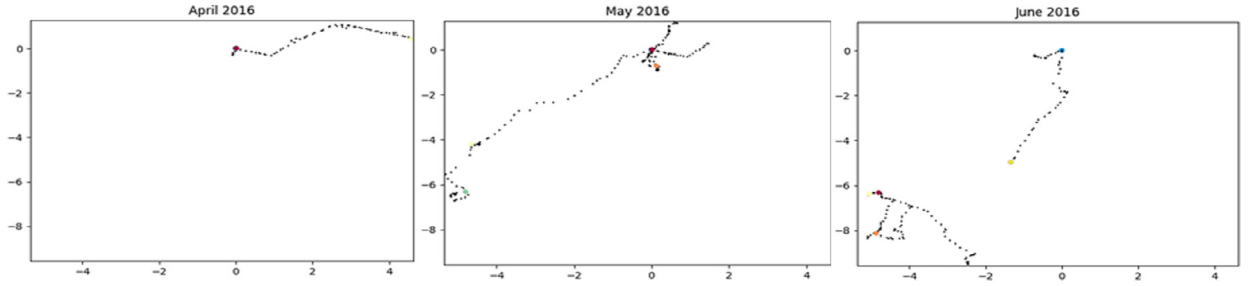

Fig. 11. Monthly clusters of a device with 3 months of data. The plots indicate significant change in movement patterns over time.

3.4. Detection of unusual movements

The data can be prepared for detection of patterns in unusual activities, including wandering patterns. Note that in the presented project we were unable to link data to any AD information, thus could not distinguish data corresponding to wandering episodes from those that are anomalies simply not following patterns (i.e., individuals with AD accompanying family members). In order to prepare data for the detection of unusual movements, the data D^{w}_{Norm} are filtered to remove all points belonging to normal locations $C_1,...,C_K$ as well as points in between (i.e., driving or walking between clusters $C_1,...,C_K$). The resulting dataset D_{Loc} includes local trajectories originating and ending in the same cluster, some of which may be indicative of wandering events.

Supervised learning methods, such as Random Forest used previously, can be then applied to identify if there are patterns in frequency of wandering and general area in which individuals with AD go. We performed a simple experiment in which RF was applied to such data. Preliminary results are shown in Table 7 indicating that it is possible to predict which specific route a person will take. However, more work is needed to fully understand methods needed for such trajectory prediction.

Table 7
Prediction of unusual movement statistics.

AUC (% of devices)	AUC <= 0.60 (55%)		0.60 < AUC <= 0.75 (26%)		AUC > 0.75 (18%)		Total	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Accuracy	0.87	0.15	0.86	0.14	0.89	0.12	0.89	0.14
Number of days	75.65	75.87	140.61	138.79	140.96	213.43	99.54	125.39
Number of data points	4400.66	5610.24	9553.92	9887.60	6651.58	8810.53	5572.47	7390.75
Number of unusual points	68.88	132.54	161.31	211.92	132.50	207.79	104.87	175.76

4. Conclusion

The majority of people with Alzheimer's Disease are in danger of wandering including getting lost outside of home. Subsequently, these individuals may get hurt, cause extreme distress for family and caregivers, and require costly search efforts. The presented research aimed at exploring possibility of using GPS devices to track people with AD and find patterns of movement that can eventually lead to prediction of wandering. The work represents first step in a long-term project aimed at understanding how IoT platforms can increase safety of people with AD, and potentially also help clinicians in tracing progression of the disease.

The presented method is based on spatiotemporal clustering used to detect normal locations where individuals typically are. Then detected clusters are passed to a supervised learning algorithm to construct models for predicting typical locations based on date and time as well as a number of features derived from geolocation data. The method achieved the best AUC (and accuracy) of 78% in predicting locations based on day, time, and extracted location and duration attributes. This high accuracy indicates highly regular lifestyle of most of individuals whose data were analyzed. This is an encouraging result, because it potentially allows for analyzing non-standard patterns of movement that may correspond to wandering and getting lost. Detection of unusual movements gave the initial AUC of 0.60 that is indicative of irregular movements outside of typical locations.

One important limitation of the presented work is that data are not limited to AD patients, and in fact there is no information about the device wearers at all. Currently, the work is being extended by directed data collection from GPS trackers linked to clinical and socioeconomic information. Individuals with confirmed stage 3-6 Alzheimer's disease will be tracked for 2-3 years to collect sufficient data for prediction of wandering, and possibly linking wandering to progression of AD. On methodological side, the movement patterns are being linked to landmark data extracted from Open Street Maps. This will allow for detection of patterns related not only to coordinates and their relationships, but more importantly to what is located at a given location.

Acknowledgements

The project would not be possible without help of GTX Corp. that provided access to GPS data and valuable expertise. The authors thank Hedyeh Mobahi, Katherine Irvin-Owen, Katherine Tompkins, Beverly Middle, Nalla Dural, and Andy Carle for collaboration and input at various stages of the project. The presented work has been supported in part by the Alzheimer's & Related Diseases Research Award Fund (ARDRAF).

References

[1] Mayo Clinic, 2017. Alzheimer's caregiving: How to ask for help. Retrieved from https://www.mayoclinic.org/healthy-lifestyle/caregivers/in-depth/alzheimers-caregiver/art-20045847.
[2] Alzheimer's Association 2017. Alzheimer's and dementia caregiver center: wandering and getting lost. Retrieved from http://www.alz.org/care/alzheimers-dementia-wandering.asp#who.
[3] D.L. Algase, D.H. Moore, C. Vandeweerd, D.J. Gavin-Dreschnack, Mapping the maze of terms and definitions in dementia-related wandering, Aging Mental Health 11 (6) (2007) 686–698.
[4] M.A. Rowe, V. Bennett, A look at deaths occurring in persons with dementia lost in the community, Am. J. Alzheimer's Dis. Other Dementias® 18 (6) (2003) 343–348.
[5] N. Ali, S.L. Luther, L. Volicer, et al., Risk assessment of wandering behavior in mild dementia, Int. J. Geriatr. Psychiatry 31 (4) (2016) 367–374.
[6] Y.T. Yang, C.G. Kels, Does the shoe fit? Ethical, legal, and policy considerations of global positioning system shoes for individuals with Alzheimer's disease, J. Am. Geriatr. Soc. 64 (8) (2016) 1708–1715.
[7] N. Shoval, G.K. Auslander, T. Freytag, et al., The use of advanced tracking technologies for the analysis of mobility in Alzheimer's disease and related cognitive diseases, BMC Geriatr. 8 (1) (2008) 7.
[8] N. Shoval, H.W. Wahl, G. Auslander, et al., Use of the global positioning system to measure the out-of-home mobility of older adults with differing cognitive functioning, Ageing Soc. 31 (5) (2011) 849–869.
[9] D. Martino-Saltzman, B.B. Blasch, R.D. Morris, L.W. McNeal, Travel behavior of nursing home residents perceived as wanderers and nonwanderers, Gerontechnology 31 (5) (1991) 666–672.
[10] N.K. Vuong, S. Chan, C.T. Lau, K.M. Lau, Feasibility study of a real-time wandering detection algorithm for dementia patients, Proceedings of the First ACM MobiHoc Workshop on Pervasive Wireless Healthcare, ACM, 2011, p. 11.
[11] N.K. Vuong, S. Chan, C.T. Lau, Application of machine learning to classify dementia wandering patterns, Gerontechnology 13 (2) (2014) 294.
[12] A. Delaunay, J. Guérin, Wandering detection within an embedded system for Alzheimer suffering patients, In 2017 AAAI Spring Symposium Series, 2017.
[13] Q. Lin, D. Zhang, X. Huang, H. Ni, X. Zhou, Detecting wandering behavior based on GPS traces for elders with dementia, in: Proceedings of the 12th International Conference on Control Automation Robotics & Vision (ICARCV), IEEE, 2012, pp. 672–677.

[14] F. Sposaro, J. Danielson, G. Tyson, iWander: An Android application for dementia patients, in: Proceedings of the Annual International Conference of IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2010, pp. 3875–3878.
[15] J. Yin, Q. Yang, J.J. Pan, Sensor-based abnormal human-activity detection, IEEE Trans. Knowl. Data Eng. 20 (8) (2008 Aug) 1082–1090.
[16] W.D. Kearns, J.L. Fozard, V.O. Nams, J.D. Craighead, Wireless telesurveillance system for detecting dementia, Gerontechnology 10 (2) (2011) 90–102.
[17] J.Y. Tung, R.V. Rose, E. Gammada, Measuring life space in older adults with mild-to-moderate Alzheimer's disease using mobile phone GPS, Gerontology 60 (2) (2014) 154–162.
[18] D. Ashbrook, T. Starner, Using GPS to learn significant locations and predict movement across multiple users, Personal and Ubiquitous computing 7 (5) (2003) 275–286.
[19] Y. Zheng, Q. Li, Y. Chen, X. Xie, W.Y. Ma, Understanding mobility based on GPS data, in: Proceedings of the 10th International Conference on Ubiquitous Computing, ACM, 2008, pp. 312–321.
[20] V.W. Zheng, Y. Zheng, X. Xie, Q. Yang, Collaborative location and activity recommendations with GPS history data, in: Proceedings of the 19th International Conference on World Wide Web, ACM, 2010, pp. 1029–1038.
[21] J. Hightower, S. Consolvo, A. LaMarca, I. Smith, J. Hughes, Learning and recognizing the places we go, in: Proceedings of the International Conference on Ubiquitous Computing, Springer, Berlin, Heidelberg, 2005, pp. 159–176.
[22] M. Feher, B. Forstner, Identifying and utilizing routines of human movement, in: Proceedings of the 2nd Eastern European Regional Conference on the Engineering of Computer Based Systems (ECBS-EERC), IEEE, 2011, pp. 135–138.
[23] M. Lin, W.J. Hsu, Mining GPS data for mobility patterns: a survey, Pervasive Mob. Comput. 12 (2014) 1–6.
[24] A. Liaw, M. Wiener, Classification and regression by randomForest, R News 2 (3) (2002) 18–22.
[25] J. Sander, M. Ester, H.P. Kriegel, X. Xu, Density-based clustering in spatial databases: the algorithm GDBSCAN and its applications, Data Min. Knowl. Discov. 2 (2) (1998) 169–194.
[26] M. Ankerst, M.M. Breunig, H.P. Kriegel, J. Sander, OPTICS: ordering points to identify the clustering structure, in: Proceedings of the ACM SIGMOD Record, 28, ACM, 1999, pp. 49–60.
[27] A.M. Olteanu, K. Huguenin, R. Shokri, M. Humbert, J.P. Hubaux, Quantifying interdependent privacy risks with location data, IEEE Trans. Mob. Comput. 16 (3) (2017) 829–842.
[28] J. Lin, R.G. Cromley, Inferring the home locations of Twitter users based on the spatiotemporal clustering of Twitter data, Trans. GIS 22 (1) (2018) 82–97.
[29] S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, M. Martonosi, J. Rowland, A. Varshavsky, Identifying important places in people's lives from cellular network data, in: Proceedings of the International Conference on Pervasive Computing, Springer, Berlin, Heidelberg, 2011, pp. 133–151.
[30] J.A. Hartigan, Clustering Algorithms, John Wiley & Sons, New-York, 1975.
[31] Y. Tuo, X. Yun, Y. Zhang, Mining Users' important locations and semantics on cellular network data, in: Proceedings of the IEEE Second International Conference on Data Science in Cyberspace (DSC), IEEE, 2017, pp. 283–291.
[32] D. Ashbrook, T. Starner, Learning significant locations and predicting user movement with GPS, in: Proceedings of the Sixth International Symposium on Wearable Computers (ISWC 2002)., IEEE, 2002, pp. 101–108.
[33] M. Kryszkiewicz, Ł. Skonieczny, Faster clustering with DBSCAN, Intelligent Information Processing and Web Mining, Springer, Berlin, Heidelberg, 2005, pp. 605–614.
[34] J. Bian, D. Tian, Y. Tang, D. Tao, A survey on trajectory clustering analysis, 2018. arXiv:1802.06971.
[35] Y. Wang, K. Qin, Y. Chen, P. Zhao, Detecting anomalous trajectories and behavior patterns using hierarchical clustering from taxi GPS data, ISPRS Int. J. Geo-Inf. 7 (1) (2018) 25.
[36] L. Cai, S. Li, S. Wang, Y. Liang, GPS trajectory clustering and visualization analysis, Ann. Data Sci. 5 (1) (2018) 29–42.
[37] S. Tasnim, J. Caldas, N. Pissinou, S.S. Iyengar, Z. Ding, Semantic-aware clustering-based approach of trajectory data stream mining, in: Proceedings of the International Conference on Computing, Networking and Communications (ICNC), IEEE, 2018, pp. 88–92.
[38] C. Zhou, D. Frankowski, P. Ludford, S. Shekhar, L. Terveen, Discovering personally meaningful places: An interactive clustering approach, ACM Transactions on Information Systems (TOIS) 25 (3) (2007) 12.