Received 29 May 2016
Revised 29 May 2016
Accepted 7 July 2016



# Contextual location prediction using spatio-temporal clustering

Djamel Guessoum
*Department of Electrical Engineering, Ecole de Technologie Superieure, Montreal, Canada*
Moeiz Miraoui
*Higher Institute of Applied Science and Technology, University of Gafsa, Gafsa, Tunisia, and*
Chakib Tadj
*Department of Electrical Engineering, Ecole de Technologie Superieure, Montreal, Canada*

## Abstract

**Purpose** – The prediction of a context, especially of a user's location, is a fundamental task in the field of pervasive computing. Such predictions open up a new and rich field of proactive adaptation for context-aware applications. This study/paper aims to propose a methodology that predicts a user's location on the basis of a user's mobility history.

**Design/methodology/approach** – Contextual information is used to find the points of interest that a user visits frequently and to determine the sequence of these visits with the aid of spatial clustering, temporal segmentation and speed filtering.

**Findings** – The proposed method was tested with a real data set using several supervised classification algorithms, which yielded very interesting results.

**Originality/value** – The method uses contextual information (current position, day of the week, time and speed) that can be acquired easily and accurately with the help of common sensors such as GPS.

**Keywords** Context-awareness, Pervasive computing, Clustering, DBSCAN, Location prediction

**Paper type** Research paper

## 1. Introduction

Human behaviour is complex and usually context-dependent according to Do and Gatica-Perez (2012). Because of its diverse applications, context prediction is a very important research topic in pervasive computing. It opens up a new and rich field of proactive adaptation for context-aware applications that adapt to the changing context in which they operate. This adaptation allows for more efficient interactions with the user, resulting in the proposal of relevant information and services (Lee *et al.*, 2010) based on the existence of services related to future contexts.

Applications that predict future contexts are linked to the current context, which the application can acquire. An example of such services is the reconfiguration of a pervasive system that includes changing the system configuration according to changes in the environments for example, accident prevention, management of personal information such as SMS user alerts and email notifications for appointments and actions, scheduling actions according to predictions made by a context-aware

International Journal of Pervasive Computing and Communications
Vol. 12 No. 3, 2016
pp. 290-309

1742-7371
DOI 10.1108/IJPCC-05-2016-0027

application (Mayrhofer, 2005) and management of device resources. Studies on context models have shown that a user's location, time and activity are the most important parameters that determine the type of service to be provided (Yuan and Herbert, 2014; Bolchini *et al.*, 2007).

Here, we are interested in predicting the location as contextual information, which is the most commonly used form of context (Ashbrook and Starner, 2003). The importance of context (location) has been studied extensively (Voigtmann and David, 2012), in particular concerning the richness of information that can be exploited in a pervasive system. For example, the location "at home" implies an individual's personal state, environment and system, as well as the type of service that this individual expects at that location. This information can be acquired easily and accurately using techniques such as GPS, WLAN and Wi-Fi. Predicting the location has diverse applications, for example, assistance and suggestions for taxi drivers to find the best routes (Do and Gatica-Perez, 2012); dissemination of information related to points of interest such as advertising, recreation or notifications of events and services available in the vicinity of these points of interest (Scellato *et al.*, 2011).

In the present paper, context and contextual information are used interchangeably and identical concepts are designated in accordance with Meissen *et al.* (2004) and Kayes *et al.* (2014), who consider context to be an instantiation of all available contextual information (e.g. location, temperature) at a certain time. We also use the widely cited definition by Dey and Abowd (1999), who defines context as:

> […] any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves.

Context is represented by a data set collected from a user's physical or system environment.

We present an approach for the prediction of a user's outdoor location on the basis of current context information that we consider to be important for the prediction. This contextual information comprises the user's current location, the day of the week, the time and the user's speed of locomotion. Most studies on the prediction of outdoor locations use additional contextual information such as traffic congestion, climate and geographical data. Although this information can improve the prediction, its collection requires additional equipment, which may introduce additional uncertainty to the accuracy of the collected data and affect the actual implementation. In the present study, spatial data were temporally segregated according to the day of the week and the hour of the day. For each hour of the day, we used the spatial density clustering algorithm density-based spatial clustering of applications with noise (DBSCAN) to determine a set of close points that define a cluster and for which the user's speed of locomotion was below a certain threshold. This speed threshold defines the boundary between a user in transit (using a vehicle such as a car, bus or train) and a user visiting a region of space and moving slowly. This threshold may be the subject of future research. For the present study, it was our goal to determine the points in observational space collected by GPS (including speed, time and day) for places that a user visits frequently on particular days, at certain times and in a specific order.

Section 2 presents a literature review and our motivations for the present study and its potential contributions. Section 3 describes the most commonly used prediction

techniques, and in Section 4, we approach clustering in general and spatial clustering in particular. Section 5 explains our application in detail. Finally, our conclusions are given in Section 6.

## 2. Related work

In the literature, context (location) prediction is considered an important aspect of improving the performance of context-aware applications to provide appropriate services. Context prediction studies can be categorised according to the type of context to predict an individual's future location, next activity, etc. The contexts that can be exploited to accomplish this prediction (such as traffic congestion or climate), if the context to predict is location, whether it is indoor or outdoor, and lastly, the type of prediction algorithm used (Bayesian networks, Markov chains, etc.).

Although the prediction algorithms are well known, the contexts to predict and the contexts that can be exploited for prediction vary and depend on the application of the prediction in each case.

In the field of context prediction, contextual information has been exploited to improve prediction performance. The types of information most used in the literature are the user's profile (e.g. senior, tourist, taxi driver), location and time (Bar-David and Last, 2016). The location of a user provides the most common and important contextual information and is the easiest to collect through several techniques (GPS, WLAN, Wi-Fi, etc.) (Voigtmann and David, 2012).

Research on location predictions is abundant because it generally involves information that can be used to proactively provide relevant services (e.g. a user in a train station will need the train routes, time schedule, etc.).

Previous research has used traffic congestion, climate, day-of-the-week, user speed and current location data in conjunction with k-means clustering to identify interesting points, which were labelled with an external Web-based service (Bar-David and Last, 2016). This contextual information was exploited to improve the prediction of future locations. In addition to predicting the location, other information was also predicted, for example, the duration of a user's stay in an actual location or at a destination which was derived by extracting user mobility patterns in conjunction with contextual information (current location and time) (Do and Gatica-Perez, 2012; Scellato *et al.*, 2011). Other studies have simply investigated the user's destination or vehicle (Patterson *et al.*, 2003; Krumm and Horvitz, 2007; Yoon and Lee, 2008).

Contextual information has also been used to improve prediction robustness and to make predictions less sensitive to sensor errors (Knig *et al.*, 2013) by using the multi-context prediction approach proposed by Sigg *et al.* (2010). Another study predicted the path of a hurricane by dividing its path into two trajectories: a spatiotemporal trajectory and a contextual trajectory derived from the geographic information of traversed regions modelled as a succession of labelled polygonal cells (Buchin *et al.*, 2014).

Because the interior of a building cannot be accessed with a GPS, special equipment such as Wi-Fi or WLAN is required to predict indoor contexts and locations, for example, an activity or movement occurring inside of a building. The CRAFFT project aims to predict activities occurring in a smart home (Nazerfard and Cook, 2015), and the Smart Doorplate Project uses the Augsburg indoor location tracking benchmarks (Petzold *et al.*, 2005) to predict the movements of an employee in an office or adjacent

room (Petzold *et al.*, 2004). In these examples of indoor location prediction, the prediction algorithms are essentially classifiers that use a user's mobility history as training data. Our work did not address the prediction of indoor locations.

The prediction of the context or location in outdoor spaces which is the focus of the present study is the type of prediction most often investigated because of the growing availability of GPS sensors with increased accuracy. A history of past observations serves as the basis for context prediction and localisation studies and for the prediction of general future behaviour of an individual in an outdoor space (Do and Gatica-Perez, 2012). These methods are based on the deterministic assumption that future events are determined by past events and that whenever a situation is observed, the subsequent situation will be similar to previously observed behaviour. Daily and weekly routines are assumed to be well established, and the activities of individuals are characterised by a degree of regularity and predictability (Bar-David and Last, 2016; Scellato *et al.*, 2011). The behavioural history of an individual is modelled as a set of patterns of visited points in space and time (points of interest, activity points, stay points and stay durations). These patterns are then exploited through various prediction techniques.

Clustering (such as DBSCAN and k-means) and spatial segmentation in the form of cells are common techniques used to determine spatial and temporal patterns. Schougaard (2007) divided space into 200-m-wide cells to predict which cell a driver will enter next; the study provided corresponding probabilities based on the vehicle's current direction and the road layout. Eagle *et al.* (2009) used dynamic Bayesian networks with segmentation to predict the trajectory of signals emitted by telecommunication towers. Krumm and Horvitz (2006) used Bayesian networks to predict the next destination of a taxi driver's trip in progress by using a history of the driver's habits and destinations. The probabilities were determined based on the number of times the destinations were visited in the past, and the map was divided into cells. By taking the time and the change of direction as constraints, the authors were able to segment the trajectories and identify stop points and activities with the CDBSCAN (DBSCAN clustering with constraints) algorithm, as well as through support vector machines (SVM) with three attributes (stop duration, mean distance to the centroid of a cluster and the shortest distance between the current location and the home or workplace) (Gong *et al.*, 2015). The study did not consider points of interest at which the individual was in motion (in a park, in the city, etc.). Mobile phone network cells were used to represent the space (Anagnostopoulos *et al.*, 2009). Algorithms from the field of machine learning and the transitions between the cells were applied to predict the future location. The minimum bounding box (MBB) was used to delimit areas according to time and related contextual information (Bar-David and Last, 2016).

In all of these studies, the spatial patterns that form the mobility history were segments of space (such as telecommunication cells, polygonal cells and rectangular cells). These studies investigated areas that contained interesting features or additional information concerning user habits.

Because prediction algorithms are application-dependent, none of the prediction algorithms is superior to the others. Each algorithm has advantages and disadvantages. With respect to Bayesian networks, dynamic Bayesian networks are most frequently used because they are suitable for time series, which is the typical form of data representing a user's contextual history (Do and Gatica-Perez, 2012; Patterson *et al.*, 2003; van Kasteren and Krose, 2007). Ensemble classifiers, also known as a committee of

classifiers, represent a combination of individual classifiers and are used to reduce prediction errors. They have demonstrated accurate performance (Lee and Cho, 2010). Markov chains have proven relevant for cases in which the points of interest represent nodes and the transition between two nodes represents the probability of traveling between two points (Ashbrook and Starner, 2003).

The prediction approach presented here uses a minimal amount of contextual information (location, day of the week, time and speed) (Figure 1). Conventional sensors such as GPS, WLAN and Wi-Fi can then be used to acquire the necessary data with considerable precision. These sensors are further able to determine locations considered to be noise, which are either transitory locations (with not enough density to create a cluster or a point of interest) or points at which the user is moving at high speed.

The proposed method is simple, and collected GPS points can be used to identify and predict the points of interest. In addition, we applied a temporal segmentation to the GPS data for which the day was divided into hours. Clusters and points of interest were searched for each hour of the day. This temporal segmentation is simpler than the usual spatial segmentation mentioned in the literature. Our study shows that filtering of high-speed and noise points can improve prediction accuracy.

## 3. Context prediction techniques

The best-known prediction techniques (Boytsov and Zaslavsky, 2010) are classification algorithms from the field of machine learning, which are used to categorise unknown data (future context or location) into discrete classes (Bayesian networks, decision trees, KNN, SVM, neural networks, etc.). Other fields from which techniques have been borrowed include computational biology (alignment), data compression (Active LeZi) and branch prediction of microprocessors (state predictors).

### *3.1 Bayesian networks*

Bayes nets, naive Bayes and dynamic Bayesian networks are statistical classification algorithms based on Bayes' theorem. Naive Bayes' networks assume that the effect of an attribute's value on the value of the class to predict is independent of the value of other attributes. Dynamic Bayesian networks are an extension of Bayesian networks. They model a dynamic system that changes over time and can represent the temporal properties of contextual information. Bayesian networks are suitable for the generation of predictive models in the real world because they take into account the uncertainty inherent in all facets of human activities (Zaguia *et al.*, 2015); however, they require more data for learning.


**Figure 1.**
Contextual information for location prediction

*3.2 Markov chains for context predictions*

Markov chains are a variant of dynamic Bayesian networks. They are used to model a system with a finite number of non-overlapping states by calculating the probability of a transition from one state to another. A user's habits can be inferred from the user's past sequence of actions. Initially, the probabilities of transitions are not known. They are used primarily to address the problem of short-term location (Bar-David and Last, 2016).

*3.3 Expert systems and decision trees*

These methods are based on expert systems and rule-based engines. The aim of this approach is to build rules for prediction, which provide a clear overview of the entire system. A decision tree (e.g. classifier C 4.5) is constructed using a function (information gain ratio) that determines the structure of the tree, as well as the informative contribution of each branch. These systems are easy to understand and able to handle non-linear interactions between variables. They are not affected by outliers and can process large amounts of categorical and numerical data (Boytsov and Zaslavsky, 2010).

*3.4 Ensemble classifiers*

Also known as a committee of classifiers, ensemble classifiers can improve context prediction performance by exploiting the advantages of individual classifiers for portions of the data (Lee and Cho, 2010). There are several ways to combine individual classifiers, of which the most popular are (Anagnostopoulos *et al.*, 2009):

- *Voting*: In which each classifier votes for or predicts a class, and the selected final class is the one that received the most votes.
- *Bagging*: In which the learning data are divided randomly into several parts, each individual classifier is applied to a portion of the data and the final selected class is the one that receives the most votes.
- *Boosting*: Which uses an incremental classification involving the classification of instances that have not been classified in the previous iteration. The final selected class is determined by a weighted vote of individual classifiers for which the weights are determined based on the performance of these classifiers.

*3.5 Alignment (sequence prediction)*

This method is used if the context can be decomposed into a sequence of events. Alignment is a context prediction algorithm applied to a time series and inspired by algorithms used in computational biology. This algorithm compares two sequences of contexts (Voigtmann, 2014). A matrix, which contains penalties, uses columns to represent the context history and lines to represent the predicted contexts. This technique presents the values of past contextual information in the form of sequences that represent the context history. Each entry receives a timestamp (Knig *et al.*, 2013). First, the algorithm combines the most recently obtained values of a context into a pattern that is aligned with the sequences of the history. This alignment then returns every match for which similar patterns are found in the history that share the quality of the match. Finally, the entry that conforms to the pattern of the match in the history is considered the prediction (Gopalratnam and Cook, 2007).

*3.6 Active LeZi*

The active LeZi context prediction algorithm has been presented by (Gopalratnam and Cook, 2004; Cook *et al.*, 2003; Gopalratnam and Cook, 2007). It is based on the LZ78 data compression algorithm by Jacob Ziv and Abraham Lempel. LeZi exploits information in a user's context history by using a sliding window to form a "LeZi" trie and to calculate the probability of each possible context transition. The maximum depth of the trie corresponds to the length of the context in the user's history (Voigtmann, 2014). To predict the context, the generated trie receives the current pattern Cp and calculates the probability of all possible contexts that may follow the given context. The context with the highest probability is predicted.

*3.7 State predictor*

The state predictor approach was developed by Petzold *et al.* (2003a, 2003b). It is based on the branch prediction techniques for microprocessors that have been applied to context prediction. The state predictor is a set of patterns of an individual's context or mobility history. It can assume two states: "strong" if the person performed the same sequence twice and "weak" otherwise. A one-state predictor signifies that if a person repeats an action once this habit is predicted and there is no weak or strong habit.

The two-state predictor usually applies to repeating the same action twice, which introduces the concept of weak and strong habits.

These location prediction techniques clearly build on a user's mobility patterns inferred from the user's history. User habits are represented by a sequence of events or contexts that serve as patterns on which to base the prediction algorithms. Previously visited locations, or points that the user has shown interest in, represent an essential characteristic of these patterns and are used to predict the location. But all these techniques require a data preparation/training phase that discovers patterns of displacement based solely on the time factor to determine the sequence of the user's movements and the probabilities of transitions (Bayes, Markov) between locations and other contextual information are rarely used.

Our approach is based on clustering, which is an unsupervised classification technique for discovering similar structures in the data. It is used to determine the points of interest (PI). This technique has been applied widely in the area of predictions (Ameyed *et al.*, 2015) and has been used in several studies aimed at predicting location (Ying *et al.*, 2011; Morzy, 2007; Yavas *et al.*, 2005). These PIs are linked to contextual information (Weekday, time, position, speed) with which the clusters will be determined, as well as the predicted location of the user (Next PI) (Figure 10).

## 4. Clustering

Clustering is an unsupervised classification algorithm. It is one of the most frequently used techniques in the fields of data mining and knowledge discovery. Clustering consists of a process that groups similar unlabelled data (such as binary, categorical, numerical, interval, ordinal, relational, textual, spatial, temporal, spatiotemporal, pictorial or multimedia data) in the same cluster and assigns dissimilar data to other clusters. The main categories of clustering are partitioning methods, hierarchical methods, density-based methods, grid-based methods and model-based methods. Partitioning methods create k partitions of data from a set of n data ($k \leq n$). K-means and k-medoids are the most well-known algorithms in this category (Warren Liao, 2005).

Hierarchical clustering groups the data into trees of clusters. The two methods in this category are the divisive and the agglomerative method, which differ in their tree construction approach. The divisive method initially groups data into partial clusters, which are then grouped into a global cluster; the agglomerative method does the inverse (CHAMELEON and CURE).

Density-based clustering methods define clusters according to their data density, which depends on the proximity parameters and on the amount of data. These clustering methods are routinely used in spatial or spatiotemporal clustering. DBSCAN and OPTICS are two of the most well-known algorithms in this category. Grid-based clustering methods such as the STING approach divide the data space into cells in which clustering is performed. Examples of clustering applications include information extraction, text search, applications for spatial databases, Web applications and DNA analysis in computational biology (Berkhin, 2006).

Because a location is defined by spatial data, we chose to use the spatial clustering (DBSCAN) method to determine the locations or points of interest visited in the mobility history of an individual.

*4.1 Density-based spatial clustering of applications with noise and temporal filtering*

The trajectories of an object recorded over a period of time can provide information on a user's mobility habits, such as points of interest visited, the duration of each visit, the succession of these visits and their context dependence. A trajectory describes the behaviour of a moving object (Kisilevich *et al.*, 2009). Therefore, clustering can be used to detect groups of objects that have the same behaviour or to detect interesting behaviours in these movements (e.g. a stop point or a frequently visited point).

DBSCAN is a data clustering algorithm proposed by Ester *et al.* (1996). It uses a distance threshold (epsilon) between the data points and a minimum number of these points to discover dense regions in the data space. It is designed to discover clusters of arbitrary shapes and can detect points considered noise as well as points that cannot be classified into any cluster (Figure 2).

In Figure 2, points 1 and 2 are density reachable; points 1 and 3 are density connected; points 1, 2 and 3 are core points; point 4 is a border point; and the N points are noise points (min. points = 4 and epsilon = e). To measure the distance between two points, DBSCAN uses a distance function (e.g. Manhattan or Euclidean). A point belongs to a cluster if there is at least one other point at a distance that is less than or equal to the threshold epsilon. A point is considered noise if its distance to all other points exceeds the threshold epsilon.

Spatial clustering has several weaknesses. It is possible that two points or observations belonging to the same cluster are observed at different times, such as when

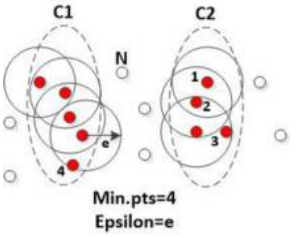


**Figure 2.**
The DBSCAN algorithm

a user passes the same point in the morning and in the evening (Figure 3). Second, the same point may be passed several times at high speed.

To resolve these two issues, we opted for a temporal segmentation of the observed points: clusters were searched at every hour of the day. The speed threshold was fixed (8 km/h) to establish an intermediate value between low and high speed (users who walk and those using vehicles), thereby enabling the extraction of high-speed points from these clusters.

In Figure 3, Points (X1; t1), (X2; t2) and (X3; t3) form the "candidate" Cluster C because they belong to the same 1-h time window (afternoon). The point (Y1; t6) cannot be included in the Cluster C because it belongs to a different time window (morning).

## 5. Experimentation

Our approach for the prediction of the location of an individual is based on the assumption that people do not move randomly and follow repetitive trajectories (Bar-David and Last, 2016). Therefore, an individual's mobility history is shaped by contextual information (day of the week, time, current location and speed). This fact is taken into account for the classification; it is represented by an $(m + 1) -$ dimensional vector v, where m is the visited locations ordered according to time, and the context represents the classifier training data. The localisation is the class of the next location to be visited (the next point of interest).

To create a user's mobility history, we used the Mobile Data Challenge (MDC) database made available by the Idiap Research Institute, Switzerland, and owned by Nokia (Kiukkonen *et al.*, 2010; Laurila *et al.*, 2012). This public data set was released as part of the Nokia MDC in 2012. The data set was collected in Switzerland from 2009 to 2011 using Nokia N95 smartphones. Although the original data set was collected with 200 participants, public data has only been released for 38 participants. The data set contains continuously collected mobility data (GPS, Wi-Fi and GSM), social interactions (voice calls, SMS and Bluetooth) and phone usage (application/data usage) for all participants. The present analysis considered mobility data only.

GSM information was scanned every minute, Wi-Fi scanning was performed every 2 minute and GPS coordinates were sampled every 10 seconds. The data set showed considerable spatial diversity because of the long collection duration and the large number of participants. The data set contains about 122 days of GPS data, 191 days of GSM data and 188 days of Wi-Fi data for nearly half of the participants. Figure 4 shows the overall user trajectories.

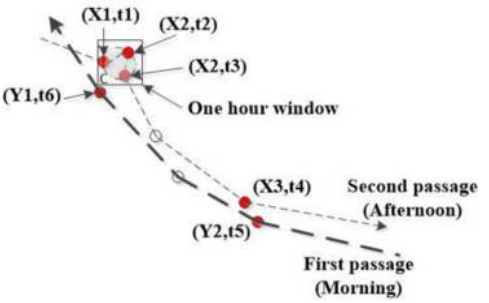


**Figure 3.**
DBSCAN and temporal filtering

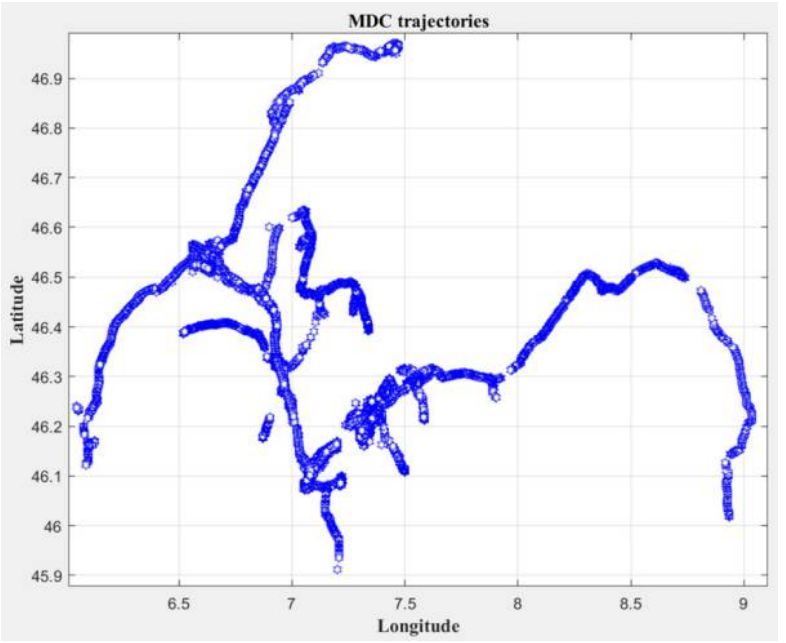


**Figure 4.**
User trajectories

For the location prediction process, we selected the GPS data (latitude, longitude, day, time, speed) of user 5,542 because this user had a rich history of visits that extended over a period of more than one year (from 2009-09-02 to 2011-02-23), as shown in Figure 5. The same process can be applied to all other users in the dataset.

*5.1 Definition of a point of interest/transition point*

*5.1.1 Point of interest.* A region of space is considered a PI (point of interest) if a user spends a significant amount of time at this location on a regular basis. To match our PI criteria for the DBSCAN, a significant amount of time was defined as a number of user observations greater than the minimum sample parameter (Figure 7):

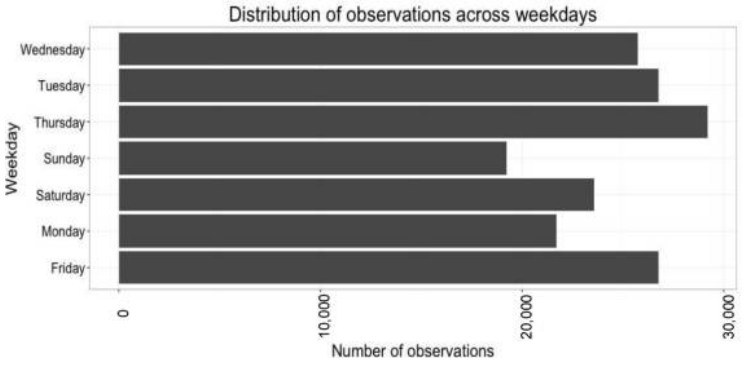


**Figure 5.**
Collected points for user #5542

- *MinSample*: Number of minimum core samples in a cluster.
- *Epsilon*: Maximum distance between points in a core sample.
- *Distance metric*: Haversine distance.

*5.1.2 Transition point.* A transition point is any point that does not belong to any cluster or PI and is considered noise. There are three types of noise points (Figure 7):

(1) *Simple legal noise*: A point that is too far away from any constructed clusters.

(2) *High-speed noise*: A point that has high speed (>8 km/h) (Figure 6).

(3) *Noise with no clusters found*: A point is located within a day of the week/hour, but no cluster exists for that combination.

A DBSCAN cluster or a point of interest is defined as follows (Figure 7):

- Day-of-the-week cluster, cluster hour in a day (0-24).
- Core samples (sets of points that define a cluster): the minimal number of points in a cluster is 200 points. The distance between two points in the cluster is 50 m (Haversine metric) (Skovsgaard *et al.*, 2014). The critical maximal speed in a cluster $V_{crit} = 8$ km/h (if $> V_{crit}$, the point is considered as -speed noise).

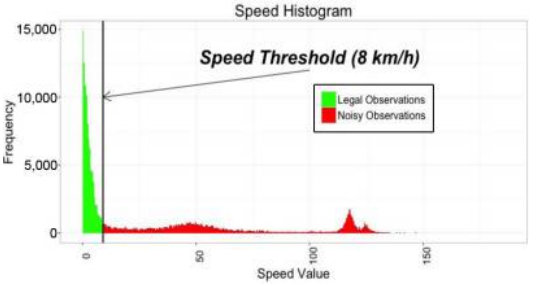


**Figure 6.**
Speed histogram

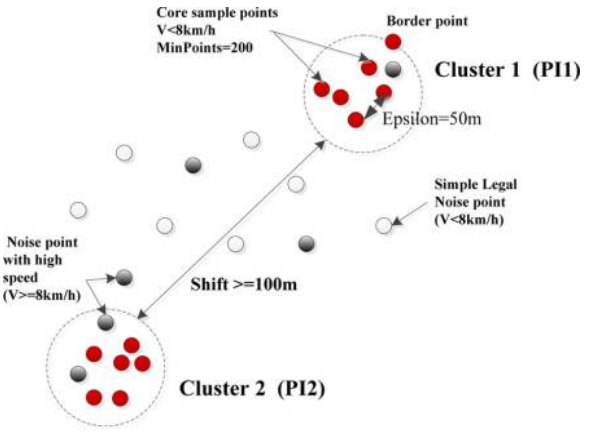


**Figure 7.**
Cluster parameters

- The cluster radius shift = 100 m, which is the distance to the nearest cluster (GPS data can vary arbitrarily by a maximum of 100 m [used during the prediction step]).
- The fill rate for high-speed noise = 0.2, which is the acceptable proportion of high-speed noise points in a cluster.

The GPS data pre-processing and DBSCAN clustering steps are shown in Figure 8.

The steps in Figure 8 are performed for the data obtained for each day. The data are prepared as follows using the programming language R (a programming language and software environment for statistical computing):

(1) Load data.

(2) Fit the DBSCAN clustering algorithm for the data of an entire day using the metric, min samples and epsilon of the configuration.

(3) Construct a set of available clusters for the current day of the week using the defined discretization, which is 1 h.
   - Examine each core sample of each cluster. If this cluster appears in the current hour, add it to this hour as an available cluster. For example, globally over all days of the week, Saturday.

Hour = 0 can include clusters [PI0, PI6], and Saturday hour = 1 can include clusters [PI0, PI3]. Therefore, this is the set of available clusters for each hour (Figure 9).

(4) Cluster filtering:
   - Reduce a set of available clusters to a set of legal clusters for each core sample of each small cluster in each hour.
   - Compute its fill rate; if it is less than the critical fill rate, this cluster is legal.

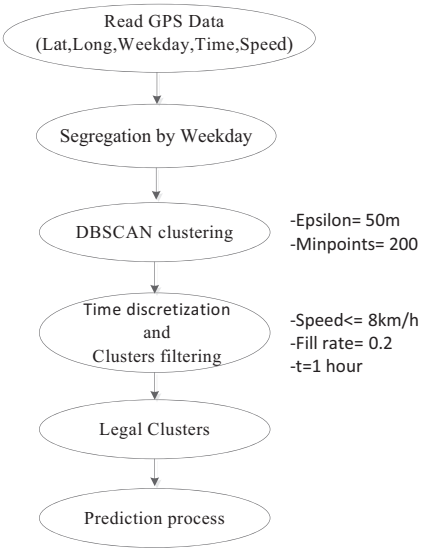


**Figure 8.**
GPS data pre-processing and DBSCAN clustering

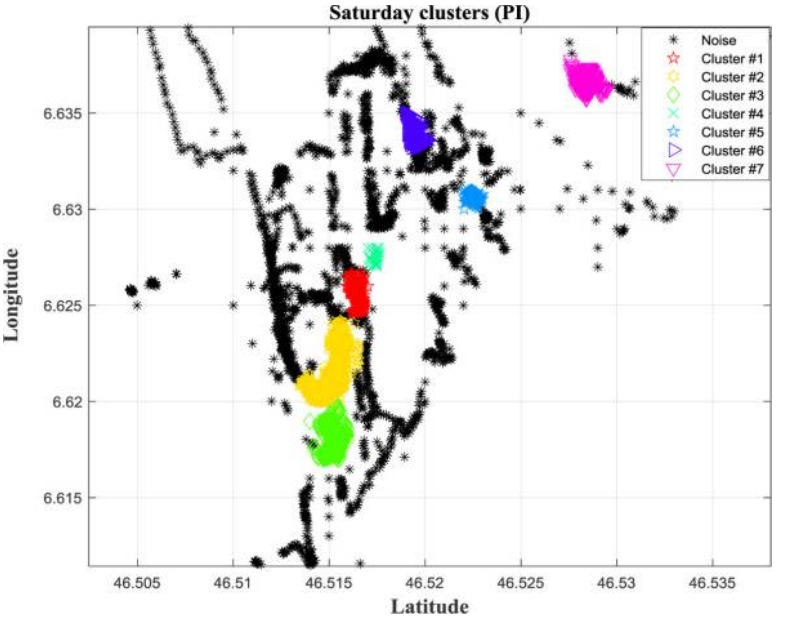


**Figure 9.**
Cluster PI Saturday

- The fill rate is computed by the algorithm.

  First, we create a subset of core samples of a cluster associated with a given hour. Then, we compute the fraction of observations whose speed is higher than the critical speed. If this fraction is greater than the fill rate, it is ignored.
- Following a filtering procedure, we can remove up to 50 of the clusters as noise clusters (the legal points of user 5,542 were generated after filtering Npoints = 172928).

*After the DBSCAN, we obtain ten clusters (seven are shown).*
*After Filtering we obtain nine clusters.*
– – – – – – – – – – – – – – – – – – – – – – – – – – – –
*ClusterId = 0, hour = 0, Center = [46.51030707 6.64090675]*
*ClusterId = 6, hour = 0, Center = [46.50975226 6.65531007]*
*ClusterId = 0, hour = 1, Center = [46.51004731 6.64004524]*
*ClusterId = 3, hour = 1, Center = [46.45 6.89]*

(5) We obtain a set of clusters for ("day", "hour").

*5.2 Prediction process*

The data obtained after pre-processing, clustering and filtering constitute the displacement history of user 5,542, modelled as a daily succession of PIs and noise. The algorithm and construction process of the predicted class "Next PI" (or the next point of interest) are given in Figures 10 and 11.

To validate the generation of the Next PI (the next point of interest) derived from the current contextual information (current location [latitude, longitude]), speed,

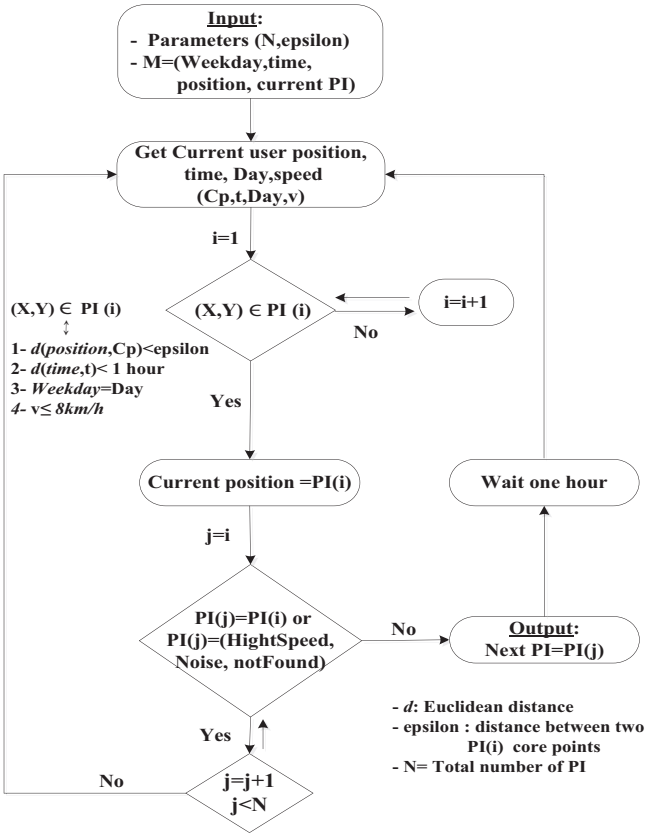


**Figure 10.**
Construction algorithm of Next PI

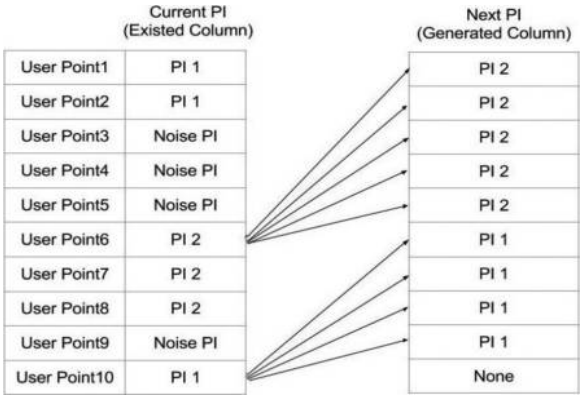


**Figure 11.**
Construction of next cluster

hour, day, PI, noise (noise, high speed), five supervised machine learning algorithms were tested using a tenfold cross validation (Naive Bayes, Bayes Nets, decision trees J48 (C4.5), KNN nearest neighbours (K = 1) and SVM. To benchmark the classification techniques, we implemented WEKA (Hall *et al.*, 2009), which is a widely used collection of machine learning algorithms for data mining software. The results are shown in Table I.

The results reported in Table I show that the location prediction methodology is effective, especially for the Bayes Net, decision trees (J48/C4.5), C 4.5, and KNN nearest neighbour classification algorithms. Additionally, the introduction of noise-prediction data improved the prediction (a mean improvement in prediction accuracy with average values of 0.654 per cent for the Bayes Net, 1.575 per cent for C4.5 and 2.102 per cent for KNN).

Graphs of the receiver operating characteristics (ROC) can be used to organise classifiers and to visualise their performance. The graphs depict the relative trade-offs between benefits (true positives) and costs (false positives). Figure 12 presents an example of ROC curves with five classifiers for one PI (point of interest #6) on Saturday 1. This is typical of other points of interest for the same day and for the other days of the week.

Evidently, all area under the curve (AUC) values are higher than 0.5, which is the boundary between a random classification and a positive classification. The prediction algorithms for J48 (C4.5), KNN and Bayes Nets provide better results than other algorithms.

ROC curves were evaluated with the AUC parameter, which indicates the probability of a classifier classifying a randomly chosen positive instance relative to a randomly chosen negative instance. In our case, the AUC indicates the probability of a selected algorithm positively classifying a PI. It is well established that the performance of a classifier with an AUC of 0.9 to 0.99 is excellent. Figure 13 shows the mean AUC for the days of the week: obtained AUC values were 0.949 for Bayes Nets, 0.964 for J48/C4.5 and 0.953 for KNN.

We tested the proposed prediction methodology with the data of a second user, user #5578. The results demonstrate that the prediction precision is similar (Table II).

Comparable accuracies were obtained for the prediction of the next point of interest of the two users (user #5542 and user #5578) and for the evaluation parameters of the classifiers (AUC), which shows that, despite the simplicity of our methodology, interesting results were obtained for the prediction of a user's future location, especially with Bayes Nets, J48/C4.5 and KNN.

## 6. Conclusion

We propose a methodology for the prediction of a user's outdoor location derived from contextual data (current location, day of the week, time and speed), which can be collected with a GPS device or with a smartphone This methodology is based on spatial clustering of data and on time segmentation to find points of interest that the user visits every day and every hour. With these points of interest, data considered noise were used to improve the prediction. It may be the subject of future work to determine whether points considered noise are relevant to improving the location prediction.

The results of this classification/prediction, which was based on two user profiles in the MDC data set, demonstrate the relevance and simplicity of our methodology. We are planning a follow-up study to introduce the semantic identification of points of interest and improve the spatial clustering process to include interesting regions that have an insufficient number of points to create a cluster.

| Day | Naïve Bayes % | | Bayes Net % | | Decision trees J48 = C4.5% | | SVM % | | Nearest neighbor Lbk (K = 1) % | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *With noise* | Without noise | *With noise* | Without noise | *With noise* | Without noise | *With noise* | Without noise | *With noise* | Without noise |
| Saturday | *29.69* | 66.62 | *77.30* | 76.43 | *92.37* | 92.11 | *69.54* | 73.19 | *93.20* | 90.36 |
| Sunday | *31.64* | 44.15 | *74.25* | 70.20 | *91.22* | 88.22 | *59.56* | 58.32 | *92.46* | 87.84 |
| Monday | *52.94* | 54.78 | *69.73* | 69.55 | *88.06* | 83.94 | *65.43* | 66.27 | *89.47* | 86.10 |
| Tuesday | *40.36* | 41.31 | *65.18* | 64.76 | *87.58* | 86.37 | *55.69* | 60.81 | *89.74* | 88.42 |
| Wednesday | *54.80* | 65.03 | *75.60* | 75.86 | *92.11* | 89.71 | *65.04* | 71.78 | *93.98* | 91.95 |
| Thursday | *31.100* | 35.28 | *61.41* | 60.42 | *84.72* | 83.63 | *54.1* | 60.17 | *86.39* | 83.75 |
| Friday | *51.26* | 50.80 | *67.96* | 69.63 | *86.72* | 87.77 | *58.99* | 58.99 | *89.89* | 89.89 |

**Table I.**
Location prediction accuracy with noise (Bold) and without noise (User #5542)

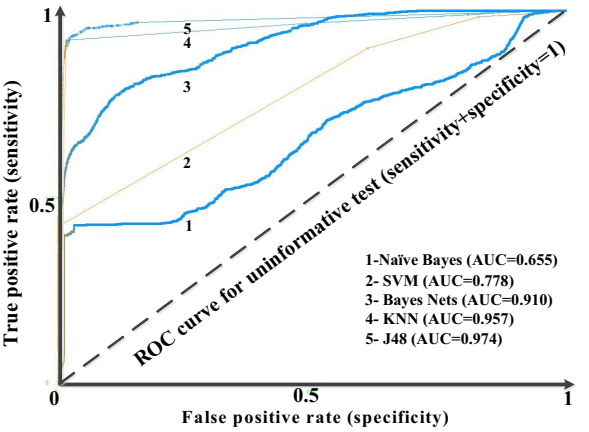


**Figure 12.**
Saturday (point of interest #6) ROC (receiver operating characteristic)

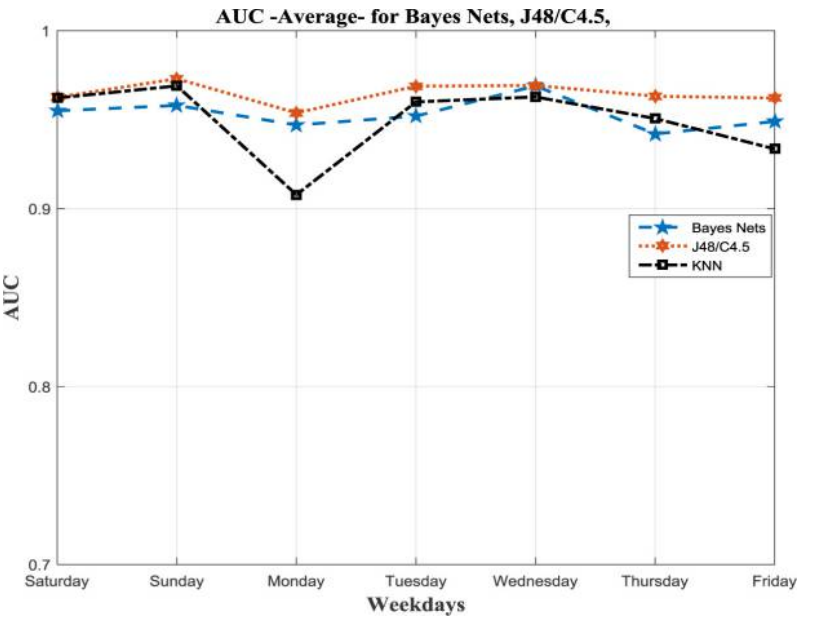


**Figure 13.**
Average day-AUC for Bayes Net, J48/C4.5, and KNN

**Table II.**
Average location prediction accuracy (User #5578)

| | Naïve Bayes % | Bayes Net % | Decision trees J48 = C4.5 % | SVM % | Nearest neighbor Lbk (K = 1) % |
|---|---|---|---|---|---|
| Prediction accuracy (average) | 67.56 | 86.54 | 97.09 | 77.48 | 98.09 |

## References

Ameyed, D., Miraoui, M. and Tadj, C. (2015), "A survey of prediction approach in pervasive computing", *International Journal of Scientific & Engineering Research*, Vol. 6 No. 5, pp. 306-316.

Anagnostopoulos, T., Anagnostopoulos, C., Hadjiefthymiades, S., Kyriakakos, M. and Kalousis, A. (2009), "Predicting the location of mobile users: a machine learning approach", *Proceedings of the 2009 International Conference on Pervasive Services*, *Paris*, pp. 65-72.

Ashbrook, D. and Starner, T. (2003), "Using gps to learn significant locations and predict movement across multiple users", *Personal and Ubiquitous Computing*, Vol. 5 No. 1, pp. 275-286.

Bar-David, R. and Last, M. (2016), "Context-aware location prediction", *Big Data Analytics in the Social and Ubiquitous Context: 5th International Workshop on Modeling Social Media, MSM 2014, 5th International Workshop on Mining Ubiquitous and Social Environments, MUSE 2014, and First International Workshop on Machine Learning for Urban Sensor Data, SenseML*, Revised Selected Papers, Springer, pp. 165-185.

Berkhin, P. (2006), "A survey of clustering data mining techniques", *Grouping Multidimensional Data*, Springer, New York, NY, pp. 25-71.

Bolchini, C., Curino, C.A., Quintarelli, E., Schreiber, F.A. and Tanca, L. (2007), "A data-oriented survey of context models", *ACM Sigmod Record*, Vol. 36 No. 4, pp. 19-26.

Boytsov, A. and Zaslavsky, A.B. (2010), "Context prediction in pervasive computing systems: achievements and challenges", in Burstein, F., Brzillon, P. and Zaslavsky, A.B. (Eds), *Supporting Real Time Decision-Making, Vol. 13 of Annals of Information Systems*, Springer, New York, NY, pp. 35-63.

Buchin, M., Dodge, S. and Speckmann, B. (2014), "Similarity of trajectories taking into account geographic context", *Journal of Spatial Information Science*, Vol. 9 No. 1, pp. 101-124.

Cook, D.J., Youngblood, G.M., Heierman, E.O. III., Gopalratnam, K., Rao, S., Litvin, A. and Khawaja, F. (2003), "Mavhome: an agent-based smart home", *In PerCom*, Vol. 3, pp. 521-524.

Dey, A.K. and Abowd, G.D. (1999), "Towards a better understanding of context and context-awareness", Technical report, GA Institute of Technology, College of Computing, Atlanta.

Do, T.M.T. and Gatica-Perez, D. (2012), "Contextual conditional models for smartphone-based human mobility prediction", *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, *ACM*, pp. 163-172.

Eagle, N., Clauset, A. and Quinn, J.A. (2009), "Location segmentation, inference and prediction for anticipatory computing", AAAI spring symposium: technosocial predictive analytics, AAAI, pp. 20-25.

Ester, M., Kriegel, H.-P., Sander, J. and Xu, X. (1996), "A density-based algorithm for discovering clusters in large spatial databases with noise", *Kdd*, Vol. 96 No. 34, pp. 226-231.

Gong, L., Sato, H., Yamamoto, T., Miwa, T. and Morikawa, T. (2015), "Identification of activity stop locations in gps trajectories by density-based clustering method combined with support vector machines", *Journal of Modern Transportation*, Vol. 23 No. 3, pp. 202-213.

Gopalratnam, K. and Cook, D.J. (2004), "Active lezi: an incremental parsing algorithm for sequential prediction", *International Journal on Artificial Intelligence Tools*, Vol. 13 No. 4, pp. 917-930.

Gopalratnam, K. and Cook, D.J. (2007), "Online sequential prediction via incremental parsing: the active lezi algorithm", *IEEE Intelligent Systems*, Vol. 22 No. 1, pp. 52-58.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I.H. (2009), "The WEKA data mining software: an update", *ACM SIGKDD Explorations Newsletter*, Vol. 11 No. 1, pp. 10-18.

Kayes, A.S.M., Han, J. and Colman, A. (2014), "PO-SAAC: a purpose-oriented situation-aware access control framework for software services", in Jarke, M., Mylopoulos, J., Quix, C., Rolland, C., Manolopoulos, Y., Mouratidis, H. and Horkoff, J. (Eds), *CAiSE, Vol. 8484 of Lecture Notes in Computer Science*, Springer, New York, NY, pp. 58-74.

Kisilevich, S., Mansmann, F., Nanni, M. and Rinzivillo, S. (2009), *Spatio-Temporal Clustering*, Springer, New York, NY.

Kiukkonen, N., Blom, J., Dousse, O., Gatica-Perez, D. and Laurila, J. (2010), "Towards rich mobile phone datasets: Lausanne data collection campaign", Proc. ICPS, Berlin.

Knig, I., Klein, B.N. and David, K. (2013), "On the stability of context prediction", in Mattern, F., Santini, S., Canny, J.F., Langheinrich, M. and Rekimoto, J. (Eds), *UbiComp (Adjunct Publication)*, ACM, New York, NY, pp. 471-480.

Krumm, J. and Horvitz, E. (2006), "Predestination: inferring destinations from partial trajectories", in Dourish, P. and Friday, A. (Eds), *Ubicomp, Vol. 4206 of Lecture Notes in Computer Science*, Springer, New York, NY, pp. 243-260.

Krumm, J. and Horvitz, E. (2007), "Predestination: where do you want to go today?", *IEEE Computer*, Vol. 40 No. 4, pp. 105-107.

Laurila, J.K., Gatica-Perez, D., Aad, I., Bornet, O., Do, T.M.T., Dousse, O., Eberle, J. and Miettinen, M. (2012), "The mobile data challenge: big data for mobile computing research", pervasive computing, number EPFL-CONF-192489.

Lee, K.C. and Cho, H. (2010), "Performance of ensemble classifier for location prediction task: emphasis on markov blanket perspective", *International Journal of u- and e-Service, Science and Technology*, Vol. 3 No. 3, p. 2010.

Lee, S., Lee, K.C. and Cho, H. (2010), "A dynamic Bayesian network approach to location prediction in ubiquitous computing environments", in Papasratorn, B., Lavangnananda, K., Chutimaskul, W. and Vanijja, V. (Eds), *IAIT, Vol. 114 of Communications in Computer and Information Science*, Springer, New York, NY, pp. 73-82.

Mayrhofer, R. (2005), "Context prediction based on context histories: expected benefits, issues and current state-of-the-art", *Cognitive Science Research Paper-University of Sussex Csrp*, Vol. 577 No. 31.

Meissen, U., Pfennigschmidt, S., Voisard, A. and Wahnfried, T. (2004), "Context- and situation-awareness in information logistics", *Current Trends in Database Technology-EDBT 2004 Workshops*, Springer, New York, NY, pp. 335-344.

Morzy, M. (2007), "Mining frequent trajectories of moving objects for location prediction", in Perner, P. (Ed.), *MLDM, Vol. 4571 of Lecture Notes in Computer Science*, Springer, New York, NY, pp. 667-680.

Nazerfard, E. and Cook, D.J. (2015), "Crafft: an activity prediction model based on Bayesian networks", *Journal of Ambient Intelligence and Humanized Computing*, Vol. 6 No. 2, pp. 193-205.

Patterson, D.J., Liao, L., Fox, D. and Kautz, H. (2003), "Inferring high-level behavior from low-level sensors", *In International Conference on Ubiquitous Computing* Springer, Berlin Heidelberg, pp. 73-89.

Petzold, J., Bagci, F., Trumler, W. and Ungerer, T. (2003a), "Global and local state context prediction", In Artificial Intelligence in Mobile Systems.

Petzold, J., Bagci, F., Trumler, W. and Ungerer, T. (2003b), "The state predictor method for context prediction", *Adjunct Proceedings Fifth International Conference on Ubiquitous Computing, Citeseer*, pp. 135-147.

Petzold, J., Bagci, F., Trumler, W., Ungerer, T. and Vintan, L. (2004), "Global state context prediction techniques applied to a smart office building", *In The Communication Networks and Distributed Systems Modeling and Simulation Conference*.

Petzold, J., Bagci, F., Trumler, W. and Ungerer, T. (2005), "Next location prediction within a smart office building", *Cognitive Science Research Paper-University of Sussex CSRP*, Vol. 577 No. 69.

Scellato, S., Musolesi, M., Mascolo, C., Latora, V. and Campbell, A.T. (2011), "Nextplace: a spatio-temporal prediction framework for pervasive systems", *Pervasive Computing*, Springer, New York, NY, pp. 152-169.

Schougaard, K. (2007), "Vehicular mobility prediction by Bayesian networks", *DAIMI Report Series*, Vol. 36 No. 582.

Sigg, S., Haseloff, S. and David, K. (2010), "An alignment approach for context prediction tasks in UbiComp environments", *IEEE Pervasive Computing*, Vol. 9 No. 4, pp. 90-97.

Skovsgaard, A., Sidlauskas, D. and Jensen, C.S. (2014), "A clustering approach to the discovery of points of interest from geo-tagged microblog posts", in Zaslavsky, A.B., Chrysanthis, P.K., Becker, C., Indulska, J., Mokbel, M.F., Nicklas, D. and Chow, C.Y. (Eds), *MDM (1)*, IEEE, New York, NY, pp. 178-188.

van Kasteren, T. and Krose, B. (2007), "Bayesian activity recognition in residence for elders", *Intelligent Environments, 2007, IE 07, 3rd IET International Conference on*, *IET*, pp. 209-212.

Voigtmann, C. (2014), "An algorithmic approach for collaborative-based prediction of user contexts in ubiquitous environments under consideration of legal implications", available at: http://nbn-resolving.de/urn:nbn:de:hebis:34-2014021945130

Voigtmann, C. and David, K. (2012), "A survey to location-based context prediction", in Springer (Ed.), AwareCast 2012 (Pervasive).

Warren Liao, T. (2005), "Clustering of time series data survey", *Pattern Recognition*, Vol. 38 No. 11, pp. 1857-1874.

Yavas, G., Katsaros, D., Ulusoy, O. and Manolopoulos, Y. (2005), "A data mining approach for location prediction in mobile environments", *Data and Knowledge Engineering*, Vol. 54 No. 2, pp. 121-146.

Ying, J.J.C., Lee, W.C., Weng, T.C. and Tseng, V.S. (2011), "Semantic trajectory mining for location prediction", in Cruz, I.F., Agrawal, D., Jensen, C.S., Ofek, E. and Tanin, E. (Eds), *GIS*, ACM, Chicago, pp. 34-43.

Yoon, T.B. and Lee, J.H. (2008), "Goal and path prediction based on user's moving path data", in Kim, W. and Choi, H.J. (Eds), *ICUIMC*, ACM, Chicago, pp. 475-480.

Yuan, B. and Herbert, J. (2014), "Context-aware hybrid reasoning framework for pervasive healthcare", *Personal and Ubiquitous Computing*, Vol. 18 No. 4, pp. 865-881.

Zaguia, A., Tadj, C. and Ramdane-Cherif, A. (2015), "Context-based method using Bayesian network in multimodal fission system", *International Journal of Computational Intelligence Systems*, Vol. 8 No. 6, pp. 1076-1090.

**Corresponding author**
Djamel Guessoum can be contacted at: djamel.guessoum.1@ens.etsmtl.ca