Learning Mobility Patterns from GPS Data for Predicting Next Points of Interest

Aicha Karite[1] and Borja Sanz Urquijo[2]

[1] Institute of Communications and Navigation (KN),
German Aerospace Center (DLR),
Wessling, Germany,
aicha.karite@dlr.de
[2] Department of Computer, Electronic and Communication Technologies,
Faculty of Engineering, University of Deusto,
48014 Bilbao, Spain

Abstract. Understanding human mobility patterns is key to developing intelligent, context-aware services. This study analyzes GPS data from users' daily activities to identify personal POIs, frequently visited places such as home, workplace, or leisure locations. Using spatial-temporal clustering, we detect these POIs and model users' mobility behavior. We then apply predictive modeling techniques to forecast the next likely POI a user will visit. The objective is to enable proactive, data-driven mobility assistance that supports personalized and efficient daily movement planning. The dataset includes GPS trajectories from 288 users collected over three months (March, April, and May), and the proposed model achieves a prediction accuracy of 90%.

Keywords: Clustering algorithm; POIs; Prediction; Markov chain; Random Forest; DBSCAN; HDBSCAN.

1 INTRODUCTION

With the growing ubiquity of smartphones and wearable devices equipped with Global Positioning System (GPS) sensors [1], collecting large-scale location and mobility data has become increasingly convenient. Modern smartphones continuously record GPS coordinates while users interact with various applications, generating rich spatial and temporal data that reflect their daily movements and behaviors. Analyzing these data enables researchers to better understand human mobility patterns and extract meaningful insights about users' routines and preferences.

GPS trajectories, which consist of sequences of latitude, longitude, and timestamp, can be used to identify Points of Interest (POIs). Identifying such POIs provides valuable information for applications in urban planning, intelligent transportation systems, and personalized recommendation services. For example, POIs can help model travel demand, analyze transportation modes, and optimize city infrastructure or route planning.

Beyond identifying visited locations, predicting an individual's next location or next POI has become a major research focus. Anticipating where a user is likely to go next can enable more context-aware and efficient services, such as improved navigation, mobility assistance, and emergency response systems. Predictive mobility analysis also holds potential for public safety applications, such as locating missing persons [2], or for supporting individuals with memory impairments by notifying relatives about their whereabouts.

Previous studies have shown that people tend to follow repetitive mobility patterns over time, allowing for the modeling and prediction of their future movements [4, 9]. Various methods have been developed to extract POIs from GPS data, generally categorized into density-based and time-density-based approaches. Density-based algorithms, such as those using kernel density estimation [3], K-means [5], or DJ-Cluster [6], group nearby points into clusters representing frequent locations. Dynamic extensions of DBSCAN have also been proposed to account for movement speed and spatial variation. In contrast, time-density-based approaches consider both distance and duration between points to detect stays, as seen in [7, 8]. However, many of these methods rely on manually chosen parameters or neglect contextual factors such as transportation mode, which can impact clustering accuracy.

For next-location prediction, traditional probabilistic models such as Markov chains [10] and Hidden Markov Models (HMMs) [14, 15] have been widely used to estimate the next POI based on previous locations. More advanced versions, like the extended mobility Markov model [16], achieve improved accuracy by incorporating additional context. Nevertheless, these models often face limitations when dealing with long-term dependencies or large-scale datasets.

In recent years, Machine Learning (ML) approaches have emerged as powerful alternatives for mobility prediction. Techniques such as K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Decision Trees (DT), and Artificial Neural Networks (ANN) can capture complex nonlinear relationships within large GPS datasets [11–13]. Hybrid models that combine clustering and learning-based prediction have further enhanced accuracy and robustness [18–21].

Despite substantial progress, most existing studies focus exclusively on either POI extraction or next-location prediction, often treating them as independent tasks. In this work, we present an integrated framework that jointly identifies personalized POIs and predicts users' next likely POI using spatial–temporal mobility patterns. Unlike prior studies that rely solely on spatial clustering, our approach incorporates both spatial density and temporal regularity to better distinguish meaningful personal locations such as home, workplace, and leisure sites. By explicitly modeling the temporal dynamics of user movement and transitions between identified POIs, the framework captures individualized behavioral routines rather than population-level trends.

Another novelty of this study lies in the scale and richness of the dataset, which contains GPS trajectories from 288 users collected continuously over a three-month period (March–May). This extended temporal coverage allows for the identification of stable, recurring movement behaviors rather than short-term fluctuations. Furthermore, the predictive component of the model leverages these personalized POI patterns to achieve a next-location prediction accuracy of 90%, demonstrating robust generalization to real-world mobility data.

Overall, the proposed framework contributes to the growing field of human mobility analysis by combining POI detection and prediction within a single, data-driven system. Its emphasis on personalization, temporal awareness, and seamless integration between clustering and prediction offers a novel pathway toward proactive, context-aware mobility assistance and intelligent transportation applications.

2 System Overview

This section presents the general workflow followed to extract POIs and predict the next likely destinations for users based on GPS trajectory data. The process consists of two main stages: data preparation and the overall processing pipeline.

2.1 Data Preparation

The dataset used in this study comprises GPS trajectories from 288 users, collected between March and May, primarily within Spain. Each record includes a timestamp, an anonymous user identifier, and geographic coordinates (latitude and longitude). To protect participant privacy, all data were collected anonymously without any personally identifiable information such as names, demographic attributes, or device identifiers. Only the minimal metadata required for mobility analysis was retained. Approximately 94% of all recorded trajectories were located within Spain as we see in Fig. 1 , while the remaining 6% correspond to movements distributed across other regions worldwide.

For individualized mobility analysis, the data were grouped by user, enabling personalized POI extraction and next-location prediction while maintaining strict privacy protection.

As the dataset originates from real-world smartphone sensors, it naturally contains measurement noise, GPS drift, and irregular sampling intervals. To improve data quality and reliability, a preprocessing procedure was applied, including:

- Removing missing or duplicated records to ensure temporal continuity.
- Filtering out invalid or inconsistent coordinates, such as impossible latitude/longitude values or stationary noise.
- Reconstructing temporal order, ensuring that all data points for each user follow chronological order.
- Resampling or smoothing trajectories where necessary to handle irregular sampling intervals.

These preprocessing steps allow each user's trajectory to reasonably approximate their true mobility patterns, providing a solid foundation for subsequent POI extraction and prediction.

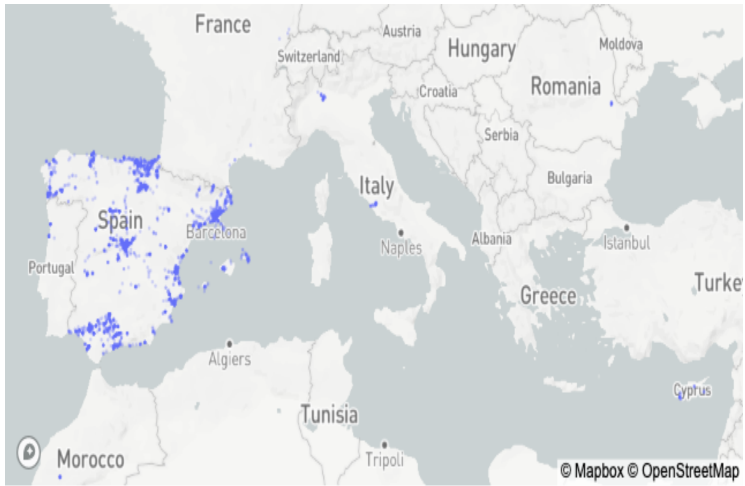

Fig. 1. Geographical Visualization of User Trajectories

2.2 Process Overview

The overall workflow of the proposed system follows a three-stage process designed to extract meaningful mobility insights from raw GPS data, as illustrated in Fig. 2. The first stage involves data preprocessing and cleaning, where noisy, incomplete, or inconsistent GPS readings are filtered to ensure high-quality input. In the second stage, Points of Interest (POIs) are identified by analyzing users' mobility traces and grouping spatially close stop points into representative locations. Finally, the third stage focuses on POI prediction, where machine learning and probabilistic models are employed to forecast each user's next likely destination based on their historical movement patterns. This structured workflow ensures a smooth transition from raw data to interpretable behavioral insights, enabling accurate mobility understanding and prediction.

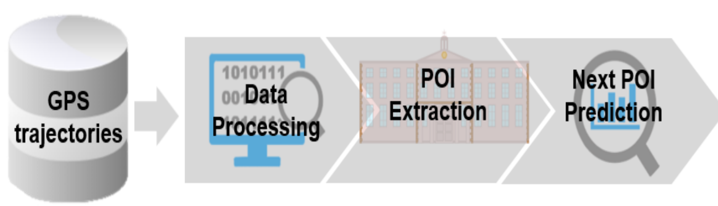

Fig. 2. System Architecture for POI Extraction and Next Location Prediction

3 Methodology

The proposed framework consists of two main stages: extraction of POIs from GPS trajectories, and prediction of the next likely POI for each user. The goal is to detect locations that users visit frequently, referred to as functional areas of daily activity, and then model the transition patterns between these locations to predict future movements.

3.1 POIs Identification

The first objective is to identify the user's POIs locations that represent meaningful activity zones. Several clustering algorithms were initially evaluated. K-Means and Gaussian Mixture Models (GMM) were found unsuitable, as they assign every data point to a cluster, making them unable to distinguish between movement paths and actual stop locations.

To handle noise and variable density in users' trajectories, we adopted a density-based clustering approach. HDBSCAN [22] performed better in detecting high-density regions corresponding to stop locations, while excluding sparse trajectory points. However, further refinement was required to accurately delineate individual POIs.

To enhance POI quality, a stop point detection algorithm [23] was implemented. This method identifies periods where a user remains within a defined spatial range for a minimum duration, indicating a stationary activity. Each stop point includes a user ID, timestamps (t_{start}, t_{end}), and coordinates.

The stop detection process depends on two parameters:

- Roaming distance (r): the maximum radius within which user movement is still considered stationary.
- Minimum stay duration (δ_t): the minimum time a user must remain within radius r for the point to be considered a stop.

To determine optimal parameter values for the stop point detection algorithm, we adopted a data driven tuning approach rather than fixing a single global configuration. Specifically, we explored a grid of r and δ_t values (e.g., $r \in \{20, 50, 100, 250, 500, 1000\}$ m and $\delta_t \in \{5, 10, 15, 20, 30, 60\}$ min) and computed diagnostic metrics for each pair, including the average number of POIs per user, time coverage of POIs. These metrics were visualized as heatmaps to identify stable regions or "plateaus," indicating parameter combinations with consistent results.

The heatmap in Fig. 3 visualizes the combined performance of Average POIs per user and Time Coverage across different parameter values for the stop-point detection algorithm. Each cell represents a pair of roaming distance (r) and minimum stay duration (δ_t), with color intensity indicating the normalized combined score. Darker colors correspond to higher combined scores, reflecting parameter combinations that detect more meaningful stops while covering a larger portion of user time. The red rectangle highlights the optimal parameters (r and δ_t) that

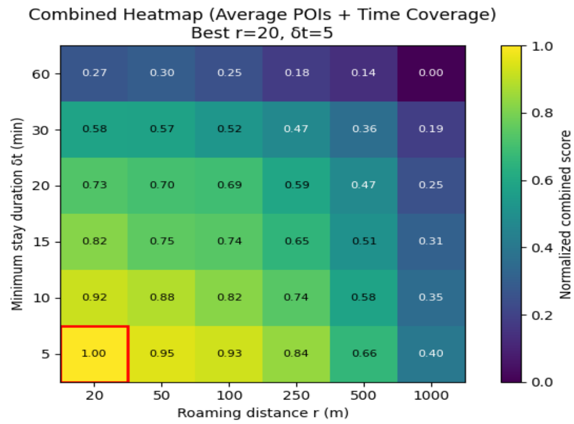

Fig. 3. Combined Heatmap of Stop-Point Detection Performance of Average POIs and Time Coverage

balance both metrics, identifying a stable region suitable for robust stop-point detection.

After extracting stop points, destinations were derived by grouping spatially close stop points together. Instead of hierarchical clustering, a KNN based approach was used. For each stop point, the KNN stops within a predefined spatial threshold were identified. Points that were mutually within each other's KNN were merged into a single group, forming a destination.

This method adapts well to varying densities of GPS data and does not require all clusters to have the same shape or size. It is also computationally efficient and robust against small positional noise, since spatial grouping is directly based on proximity relationships rather than global clustering thresholds.

The KNN grouping approach relies on two key parameters:

. Number of neighbors (k): defines how many nearby stop points are considered when forming a group.
. Maximum spatial distance (d_{max}): the maximum distance in meters between two stop points to be considered part of the same destination.

Once groups were formed, a medoid was calculated for each destination. The medoid is the stop point within a group that has the minimum average distance to all other points in the same group, representing the most central and representative coordinate of that destination.

Each destination was then assigned a visit frequency, the number of times the user visited that destination and a unique cluster identifier, providing a clear structure for subsequent prediction.

3.2 POIs Prediction

The second stage of the framework focuses on predicting each user's next likely destination based on their historical mobility behavior. After identifying all POIs, each user's trajectory was transformed into an ordered sequence of POI visits. This temporal sequence reflects how users typically transition from one destination to another throughout the day.

To train and evaluate the predictive models, the data were divided into two subsets: 80% for training and 20% for testing. The training data were used to learn users' transition and behavioral patterns, while the testing data were used to assess the generalization ability of each approach. Two distinct prediction paradigms were explored: a probabilistic sequential model and a machine learning–based model, allowing us to compare their effectiveness in modeling human mobility.

Hidden Markov Model: HMM was employed to capture the probabilistic transitions between users' destinations. Each POI is treated as a hidden state, and transitions between these states represent the likelihood of a user moving from one destination to another. The model assumes that the next location depends primarily on the current state, making it suitable for sequential and time-dependent data. The transition probabilities were estimated from the users' historical sequences of POI visits, while the emission probabilities describe the likelihood of observing specific movement events given a hidden state. The HMM captures users' regular mobility patterns and is capable of handling short-term temporal dependencies effectively. However, its performance can be limited when mobility behavior exhibits long-term dependencies or when contextual factors influence movement.

Random Forest Classifier: To address these limitations, we also implemented a Random Forest (RF) model, representing a supervised ML approach. Unlike HMMs, which rely on probabilistic state transitions, Random Forests learn complex nonlinear relationships between input features and the next predicted destination. For this purpose, several temporal and contextual features were extracted, including:

. Time of day: morning, afternoon, evening
. Day of the week: weekday vs. weekend)
. Previous visited destination (POI_{t-1})
. Visit frequency of destinations
. Travel distance and time between the last two destinations

These features were used to train the model to classify the next likely POI based on the user's past movement and temporal context. The ensemble nature of Random Forests aggregating multiple decision trees helps to minimize overfitting and enhances prediction stability.

4 Results and discussion

4.1 POIs Identification

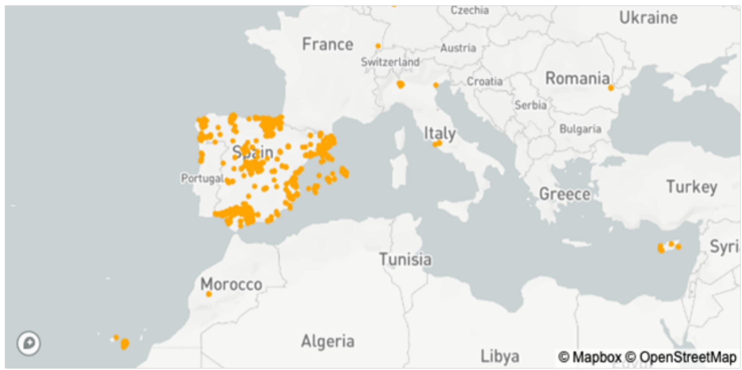

Fig. 4. Detected Stop Points Across All Users

Fig. 4 illustrates the extracted stop points for all users after applying the stop point detection algorithm. Each point on the map represents a location where a user remained stationary for a significant period, reflecting potential areas of interest. The distribution of these points demonstrates the wide geographic spread of user activities and highlights regions with dense clusters of stationary behavior. Compared to the raw trajectory data, the number of retained points was considerably reduced due to the elimination of transient GPS readings associated with movement between destinations. This reduction allows a clearer focus on meaningful stops, removing noise and redundant trajectory information.

Subsequently, nearby stop points were grouped into clusters representing distinct destinations as shown in Fig. 5. Each cluster corresponds to a frequently visited area, derived by merging spatially proximate stop points into a single representative medoid. In the visualization, different colors represent distinct destination clusters, while the size and shade of each circle indicate visit frequency, larger and lighter circles correspond to destinations visited more frequently by users, whereas darker and smaller circles represent less frequently visited places.

Overall, this two-step approach combining stop point detection and destination clustering provides a robust way to extract the user's most relevant locations, offering a cleaner and more interpretable visualization of real mobility behavior.

4.2 POIs Prediction

After identifying the destinations, two predictive approaches were evaluated to forecast each user's next likely location: a HMM and a RF Classifier. Both models

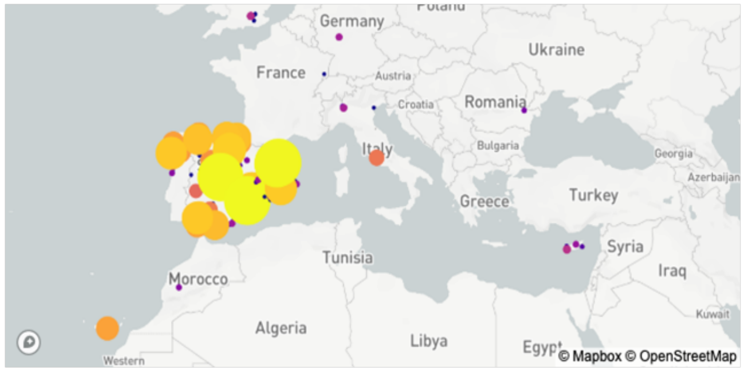

Fig. 5. Clustered Destination Points Showing Frequently Visited POIs Among Users

were trained using 80% of each user's mobility data and tested on the remaining 20%.

The HMM captures sequential dependencies between destinations by learning transition probabilities from historical visits, while the RF model integrates temporal and contextual features like time of day, day of week, last destination for classification.

Table 1. Hidden Markov Model Model Performance.

Metric	**Hidden Markov Model (HMM)**
Accuracy	73%
F1-score	72%
Macro Average	71%
Weighted Average	73%

Table 1 presents the accuracy matrix for the HMM model, and Table 2 shows the corresponding results for the RF classifier. The results indicate that the machine learning model achieved slightly higher accuracy than the probabilistic model, reflecting its ability to incorporate additional contextual variables beyond sequential dependencies.

Table 2. Random Forest Model Performance.

Metric	**Random Forest (RF)**
Accuracy	90%
F1-score	89%
Macro Average	88%
Weighted Average	90%

4.3 Discussion

The results demonstrate the effectiveness of combining stop-point detection with clustering to identify meaningful Points of Interest from noisy GPS trajectories. By grouping spatially close stop points into destinations, the system eliminates redundant trajectory data while preserving essential behavioral patterns. For the stop point detection algorithm, we optimized the parameters using a data driven approach, exploring a grid of values and selecting stable regions in the metrics to ensure robust detection across users.

In terms of prediction, both models perform reliably, but the Random Forest classifier outperforms the HMM due to its flexibility in capturing nonlinear relationships between temporal and spatial features. However, the HMM remains valuable for modeling sequential mobility patterns, particularly in scenarios where transitions between destinations follow consistent routines. The prediction models were trained separately for each user, which means that accuracy varies depending on the amount of trajectory data available; some users with limited data naturally exhibit lower predictive performance. We did not explore deep learning models due to the limited amount of data, which would be insufficient to effectively train large neural networks.

Overall, the proposed framework provides a robust pipeline for transforming raw GPS trajectories into interpretable behavioral insights. Future work will aim to integrate additional contextual data, such as transportation modes or semantic place information, to further enhance prediction accuracy and generalization.

5 Conclusions

This study presented an integrated framework for identifying POIs and predicting users' next potential locations based on GPS trajectory data. By analyzing the mobility traces of 288 users collected over a three-month period, we extracted frequently visited locations and subsequently modeled the transitions between these POIs to predict future movements.

The proposed approach achieved a prediction accuracy of 73% for the HMM and 90% for the RF model, demonstrating the superior performance of machine learning techniques in capturing individual mobility patterns and forecasting future behavior. These findings highlight the potential of GPS-based trajectory analysis for developing intelligent mobility services, including personalized route recommendations, urban planning support, and context-aware navigation systems.

Future research could focus on enhancing prediction accuracy by incorporating additional contextual information such as transportation mode, temporal behavior, or social interactions. Moreover, integrating real-time data streams and exploring advanced deep learning models could further improve adaptability and scalability in dynamic urban environments.

References

1. H. Xu and S. -B. Cho, "Recognizing Semantic Locations from Smartphone Log with Combined Machine Learning Techniques," 2014 IEEE 11th Intl Conf on Ubiquitous Intelligence and Computing and 2014 IEEE 11th Intl Conf on Autonomic and Trusted Computing and 2014 IEEE 14th Intl Conf on Scalable Computing and Communications and Its Associated Workshops, Bali, Indonesia, 2014, pp. 66-71, doi:10.1109/UIC-ATC-ScalCom.2014.128
2. Barone, Pier Matteo,Cutting-Edge Methodologies for Searching Missing Persons,Forensic Archaeology and New Multidisciplinary Approaches: Topics Discussed During the 2018-2023 European Meetings on Forensic Archaeology (EMFA), 2025, Springer Nature Switzerland, pp. 3-16, https://doi.org/10.1007/978-3-031-86308-0_1
3. Giulio Biroli and Marc Mézard, Kernel Density Estimators in Large Dimensions, 2024,arXiv, https://arxiv.org/abs/2408.05807
4. Brian P. Clarkson, Life patterns : structure from wearable sensors, 2002, https://api.semanticscholar.org/CorpusID:109356171
5. Abiodun M. Ikotun, Absalom E. Ezugwu, Laith Abualigah, Belal Abuhaija, Jia Heming, K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data, Information Sciences, Volume 622, 2023, Pages:178-210, ISSN 0020-0255, https://doi.org/10.1016/j.ins.2022.11.139
6. Jackson Bland, Sydney Miller, Elijah D. Algarin-Martinez, Abbigail M. Biggs, Maria E. D. Cavasini, Michael A. Chase, Caitlyn Coleman, Victoria Correa, Don F. Danielson, Wynter R. Dean, Joseph L. French, Mae E. Horne, Breanna M. Macumber, Francesca Kristal Martini, Sydney G. Mazzei, Coen E. E. McGarrah, Oslow Odegaard, Indu S. Parameswaran, Corrisa Quarterman, Taylor M. Rand, Kira M. Ruiz-Houston, Ava R. Sciacchitano, Nina S. Seidensticker, Anna Soltys, Adrian E. Terron-Osorio, Allyson L. Todd, Audrey R. Wood, Maxwell D. Ungrey, Richard S. Pollenz, Annotation of Secretariat and Hydrus, two DJ cluster phages isolated on Gordonia rubripertincta, Microbiology Resource Announcements, Volume 12, Issue 10, 2023, , ISSN 2576-098X, https://doi.org/10.1128/MRA.00623-23
7. hen, J., Yang, L., Huang, H., Long, X, A Task-Priority's Time-Density Based Resource Allocation Method for Resident Space Objectives Catalog Maintaining, Proceedings of 4th 2024 International Conference on Autonomous Unmanned Systems (4th ICAUS 2024), ICAUS 2024, Lecture Notes in Electrical Engineering, vol 1377 Springer, Singapore, https://doi.org/10.1007/978-981-96-3568-9_9
8. Zhandos A. Moldabekov and Sebastian Schwalbe and Uwe Hernandez Acosta and Thomas Gawne and Jan Vorberger and Michele Pavanello and Tobias Dornheim, Enhancing the Efficiency of Time-Dependent Density Functional Theory Calculations of Dynamic Response Properties, 2025,arXiv, https://arxiv.org/abs/2510.01875,
9. González, Marta, C and Hidalgo, César A, Understanding individual human mobility patterns, (2008), doi: 10.1038/nature06958.
10. J. Yang and J. Xu and M. Xu and N. Zheng and Y. Chen, Predicting next Location Using a Variable Order Markov Model, Proceedings of the 5th ACM SIGSPATIAL International Workshop on GeoStreaming, 2014, pp. 37-42, doi:10.1145/2676552.2676557.
11. Sarker, H. Iqbal, Machine Learning: Algorithms, Real-World Applications and Research Directions, SN Computer Science, 2014, pp. 2661-8907, doi:10.1007/s42979-021-00592-x.

12. Z. Zhang, Introduction to machine learning: k-nearest neighbors, Ann Transl Med, 2016 Jun, 4(11):218, doi:10.21037/atm.2016.03.37.
13. A. Hamoudzadeh and S. Behzadi, Predicting user's next location using machine learning algorithms, Spat. Inf. Res. Vol. 29, No. 3, 2021, 379-387, doi:10.1007/s41324-020-00358-2.
14. W. Mathew, R. Raposo and B. Martins, Predicting future locations with hidden Markov models, UbiComp'12: Proceedings of the 2012 ACM Conference on Ubiquitous Computing, September 2012, pp. 911-918, doi:10.1145/2370216.2370421.
15. D. Ashbrook and T. Starner, Learning significant locations and predicting user movement with GPS, Proceedings. Sixth International Symposium on Wearable Computers, 2002, pp. 101-108, doi:10.1109/ISWC.2002.1167224.
16. S. Gambs, M. O. Killijian and M. N. del Prado Cortez, Next place prediction using mobility Markov chains, MPM '12: Proceedings of the First Workshop on Measurement, Privacy, and Mobility, April 2012 Article No. 3, pp. 1-6, doi:10.1145/2181196.2181199.
17. L. Song, D. Kotz, R. Jain and Xiaoning He, Evaluating location predictors with extensive Wi-Fi mobility data, IEEE INFOCOM 2004, 2004, pp. 1414-1424 vol.2, doi:10.1109/INFCOM.2004.1357026.
18. Q. Lv, Y. Qiao, N. Ansari, J. Liu and J. Yang, Big Data Driven Hidden Markov Model Based Individual Mobility Prediction at Points of Interest, in IEEE Transactions on Vehicular Technology, vol. 66, no. 6, pp. 5204-5216, June 2017, doi: 10.1109/TVT.2016.2611654.
19. C. Ma, S. Wan, B. Han and H. Gui, A Framework for Hybrid Location Prediction via Decision Tree Classification, Proceedings of IEEE International Conference on Robotics and Biomimetics, December 12-15, 2018, Kuala Lumpur, Malaysia, doi: 10.1109/ROBIO.2018.8664721.
20. A. Noulas, S. Scellato, N. Lathia and C. Mascolo, Mining User Mobility Features for Next Place Prediction in Location-Based Services, 2012 IEEE 12th International Conference on Data Mining, 2012, pp. 1038-1043, doi:10.1109/ICDM.2012.113.
21. L. Xia, Q. Huang and D. Wu, Decision Tree-Based Contextual Location Prediction from Mobile Device Logs, Hindawi Mobile Information Systems, Vol. 2018, Article ID 1852861, 11 pages, doi:10.1155/2018/1852861.
22. Malzer, Claudia and Baum, Marcus, A Hybrid Approach To Hierarchical Density-based Cluster Selection, 2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI), pages:223–228, http://dx.doi.org/10.1109/MFI49285.2020.9235263,
23. Wang K, Pang L, Li X. Identification of Stopping Points in GPS Trajectories by Two-Step Clustering Based on DPCC with Temporal and Entropy Constraints. Sensors (Basel). 2023 Apr 5;23(7):3749. doi:10.3390/s23073749.