**TRAJECTORY DATA MINING:**

**Researching trajectory data by GPS tracks**

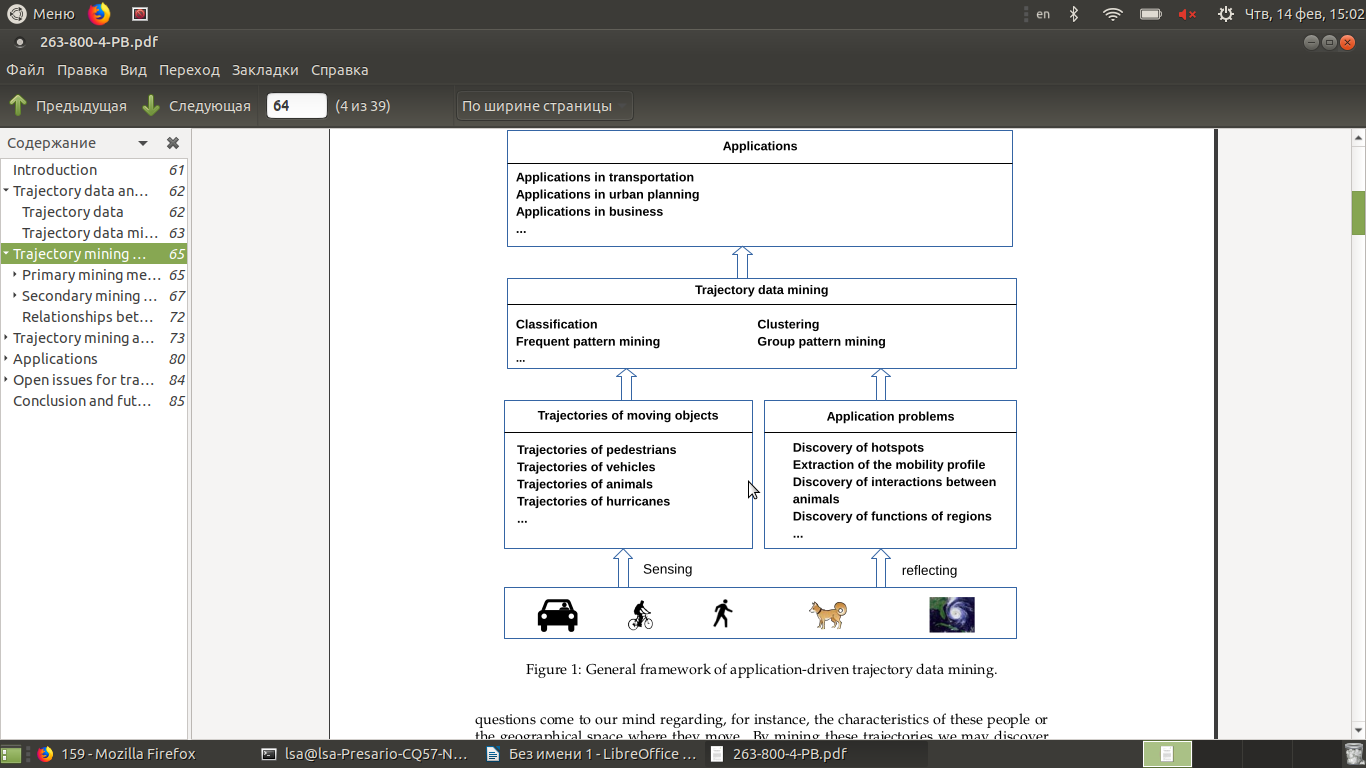
Leleko Serhiy, Palahin Volodymyr, Filipov Vitalii

**1. Define the problem**

Development of information and communication technologies, wireless communication, gives us the chance to obtain data which contain the geographical coordinates changing in time. Their analysis can lead to the solution of important research tasks in various areas, such as city planning, transport, behavioral ecology, analysis of sports trainings, observation and safety.

**Vehicles tracking (cars, taxi, freight transportation)**

Currently, many civilian vehicles, commercial and municipal transport (buses, taxis) are equipped with GPS receivers to track their trajectory [1]. On the basis of the obtained data often solves the following tasks: traffic optimization, transport movement logistics (for commercial traffic), analysis of the movement trajectory and deviations from it (match mapping) [2-3].



**Figure 1 – General Framework of application-driven trajectory data mining**

There are many scientific articles that are devoted to this subject. However, we are lack of applying data mining methods and algorithms to real data. To this end, we conduct a comprehensive survey that thoroughly explores the field of trajectory data mining (by real datasets), according to the paradigm shown in Figure 1.

In the research we will do the following [4]:

* Classify the sources generating trajectory data into some groups, listing a few key applications that trajectory data can enable in each group.
* Trajectory pre-processing: noise filtering, segmentation, and map-matching. The goal of noise filtering is to remove from a trajectory some noise points that may be caused by the poor signal of location positioning systems (e.g. when traveling in a city canyon).
* Trajectory compression is to compress the size of a trajectory (for the purpose of reducing overhead in communication, processing, and data storage).
* A stay point detection algorithm identifies the location where a moving object has stayed for a while within a certain distance threshold.
* Trajectory segmentation divides a trajectory into fragments by time interval, or spatial shape, or semantic meanings, for a further process like clustering and classification.
* Map-Matching aims to project each point of a trajectory onto a corresponding road segment where the point was truly generated.

So we can conduct next GPS-mining tasks:

* Trajectory uncertainty: Objects move continuously while their locations can only be updated at discrete times, leaving the location of a moving object between two updates uncertain. To enhance the utility of trajectories, a series of research tried to model and reduce the uncertainty of trajectories.
* Trajectory pattern mining: Long trajectories enables opportunities for analyzing the mobility patterns of moving objects, which can be represented by an individual trajectory containing a certain pattern or a group of trajectories sha-ring similar patterns.
* Trajectory classification: Using learning approaches, we can classify trajectories or segments of a trajectory into some categories, which can be activities (taxi or cargo) or different transportation modes, such as walking and driving.
* Trajectory Outlier Detection: Different from trajectory patterns that frequently occur in trajectory data, trajectory outliers (anomalies) can be items (a trajectory or a segment of trajectory) that is significantly different from other items in terms of some similarity metric.

**2. Data acquisition**

Data on the trajectories movement of the researches objects usually receive by devices equipped with GPS (or GLONASS) modules (less often with GSM / WiFi modules ). In our researches we will select those navigation devices whose work is based on GPS (GLONASS) technology as a GPS data-source, because their accuracy is quite high – up to 10 meters in open terrain, for civilian devices [5, 6]. Also we need to add, that the data received from GSM / WiFi modules will not be considered, due to their low accuracy, as for navigation devices.

**Machines and Vehicles Data**

To research the trajectories of transport (cars, buses, taxis, pedestrians), we will use the GPS trajectory dataset that was collected in (Microsoft Research Asia) Geolife project by 182 users in a period of more than five years (from April 2007 to August 2012) [7]

This dataset contains 17,621 trajectories with a total distance of 1,292,951 kilometers and a total duration of 50,176 hours. These trajectories were recorded by different GPS loggers and GPS-phones, and have a variety of sampling rates. 91.5% of the trajectories are logged in a dense representation, e.g. every 1-5 seconds or every 5-10 meters per point.

This dataset recoded a broad range of users’ outdoor movements, including not only life routines like go home and go to work but also some entertainments and sports activities, such as shopping, sightseeing, dining, hiking, and cycling.

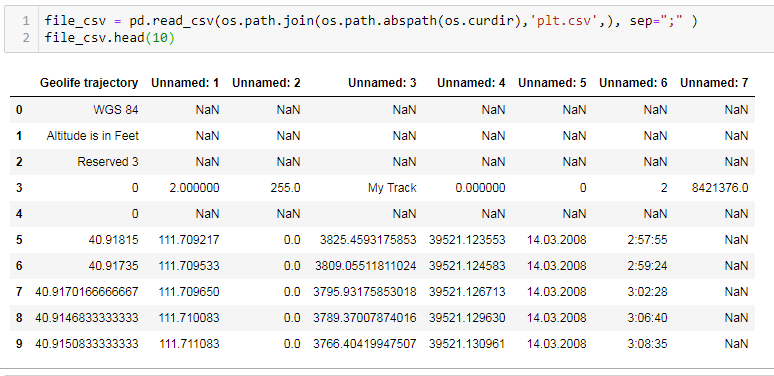
73 users have labeled their trajectories with transportation mode, such as driving, taking a bus, riding a bike and walking. There is a label file storing the transportation mode labels in each user’s folder.

Data Format and Trajectory file

Every folder of this dataset stores a user’s GPS log files, which were converted to PLT format. Each PLT file contains a single trajectory and is named by its starting time. To avoid potential confusion of time zone, we use GMT in the date/time property of each point, which is different from our previous release.

PLT format:

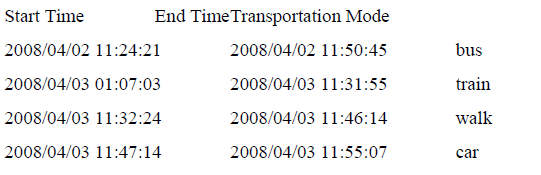
* Field 1: Latitude in decimal degrees.
* Field 2: Longitude in decimal degrees.
* Field 3: All set to 0 for this dataset.
* Field 4: Altitude in feet (-777 if not valid).
* Field 5: Date – number of days (with fractional part) that have passed since 12/30/1899.
* Field 6: Date as a String.
* Field 7: Time as a String.
* Line 1…6 are useless in this dataset, and can be ignored. Points are described in following lines, one for each line. Field 5 and field 6-7 represent the same date/time in this dataset. We may use either of them.



**Figure 1 – GPS Data in PLT format**

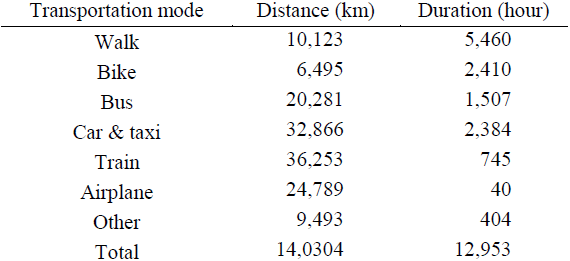
Transportation mode labels

Possible transportation modes are: walk, bike, bus, car, subway, train, airplane, boat, run and motorcycle. The date/time of all labels have converted to GMT (most of them were created in China). Example is represented on figure 2.



**Figure 2 – Transportation mode labels in GPS Data**

Full details are available for download by the link [8]. Link: <https://goo.gl/HUy1iS>

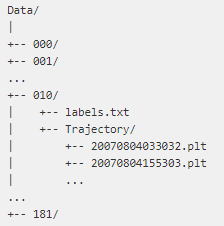


**Figure 3 – Total distance and duration of transportation modes**

The total distance and duration of transportation modes are listed in Figure 3. Though this only covers a part of the dataset used in the following papers.

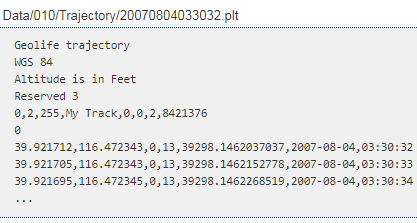
**Dataset structure**

The top level “Data/ folder” contains 182 user folders, numbered 000/ to 181/. Each user folder has a “Trajectory/ folder” containing the user’s GPS trajectories (saved as .plt files) and optionally a labels.txt file specifying the mode of transportation employed for a given time interval. Dataset folder structure are represented in Figure 4



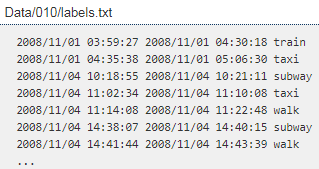
**Figure 4 –** Dataset folder structure

The .plt files starting with the seventh line, each line specifies in a comma-separated fashion the attributes of a single GPS point: (0) latitude, (1) longitude, (2) unused, (3) altitude, (4) days since Dec. 30 1899, (5) date, (6) time.



**Figure 5 –** PLT file structure

Each line in labels.txt specifies in a space-delimited fashion the start date, start time, end date, end time, and mode of transportation.



**Figure 6 –** Labels file structure

**3. Preprocessing**

**Features evaluating (Motion characteristics for GPS track**) – several motion features can be computed for every GPS point based on their geographic coordinates and timestamps.

Main features:

* The relative distance between two consecutive GPS points can be computed using the widely-used Vincenty’s formula [9]. The Vincenty’s formula is a common and accurate method for computing the geographical distance between two points on the surface of a spheroid.

RDps = Vincenty (p1[lat, lon], p2[lat, lon])

* The time interval between two successive GPS points is another motion quantity that can be simply computed.
* Having the relative distance and time interval, other fundamental kinematic motions including speed, acceleration, and jerk are computed to provide more information about a user’s motion.
* Speed is the rate of change in distance that shows how fast a user is traveling.
* Acceleration is the rate of change in speed that shows how fast a user is changing their speed.
* The rate of change in acceleration, is a significant factor in safety issues such as critical driver maneuvers.

Evaluated parameters for gps points:

* RDps – represent the relative distance;
* ∆tp1 = p2[t] − p1[t] – represent time interval;
* Sp1 = RDp1 ∆tp1 – represent speed;
* Ap1 = Sp2 − Sp1 ∆tp1 – represent acceleration/deceleration;
* Jp1 = Ap2 − Ap1 ∆tp1 – represent jerk of the point p1.

Analogously, the above-mentioned formulas are used to calculate the motion features of other GPS points. Also in the calculations we must consider the rate of change in the heading direction of different transportation modes. For example, cars and buses have to move only alongside existing streets while people with walk or bike modes alter their directions more frequently.

**Algorithm for preprocessing trajectories**

* dividing the GPS trajectory of a user’s trip into the segments with only one transportation mode.
* evaluation (representation) parameters for each GPS segment is designed.

First step of proposed algorithm for partitioning a user’s trip (preprocessing trajectories) into segments with a unique transportation mode consists of two parts:

* Considering for having all input samples with a fixed size, all GPS trajectories need to be either truncated or padded to a fixed size.
* Separation of different types of activity in one track – majority of GPS segments contains one or two transportation modes after this uniform-size segmentation, which improves the performance of the overall segmentation process.

Code: repository

**Sources of literature**

1. O'Flaherty C.A.. Transport Planning and Traffic Engineering / C.A. O'Flaherty, M.GH. Bell, P.W. BonsaU, G.R. Leake, A.D. May, C.A. Nash : Linacre House, Jordan Hill, Oxford OX2 8DP, UK. – 2006 – p. 532
2. Alt H. Matching Planar Maps / H. Alt., A. Efrat, G. Rote, C. Wenk. Journal of Algorithms. – 2003. – vol. 49:262 – p. 283.
3. Brakatsoulas S. On map-matching vehicle tracking data / S. Brakatsoulas, D. Pfoser, R. Salas, C. Wenk. In Proceedings 31st International Conference on Very Large Data Bases (VLDB). – ACM, 2005. – p 853–864.
4. Yu Zheng. Trajectory Data Mining: An Overview: Transaction on Intelligent Systems and Technology. – ACM Trans. – Vol. 6, – No. 3, Article 1. – Sept. 2015.
5. Georg Thieme Verlag. Accuracy of GPS Devices for Measuring High-intensity Running in Field-based Team Sports / E. Rampinini, G. Alberti, M. Fiorenza, M. Riggio, R. Sassi, T. O. Borges, A. J. Coutts. New York: Int J Sports Med. – 2015, – 36(01). – pp.49-53.

Access mode:

<https://www.thieme-connect.com/products/ejournals/html/10.1055/s-0034-1385866>

1. The U.S. government publishes website to educate the public about the GPS: [сайт]. – Режим доступа: <https://www.gps.gov/systems/gps/performance/accuracy/>
2. GeoLife: Building Social Networks Using Human Location History: [сайт]. – Режим доступа: <https://goo.gl/F8EUsf>
3. GeoLife GPS Trajectories: [сайт]. – Режим доступа: <https://goo.gl/HUy1iS>
4. T. Vincenty. Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations / Survey review. – vol. 23. – 176, pp. 88–93.
5. GeoLife PLT preprocessing. Access mode:

https://heremaps.github.io/pptk/tutorials/viewer/geolife.html