**TRAJECTORY DATA MINING:**

**Researching trajectory data by GPS tracks**

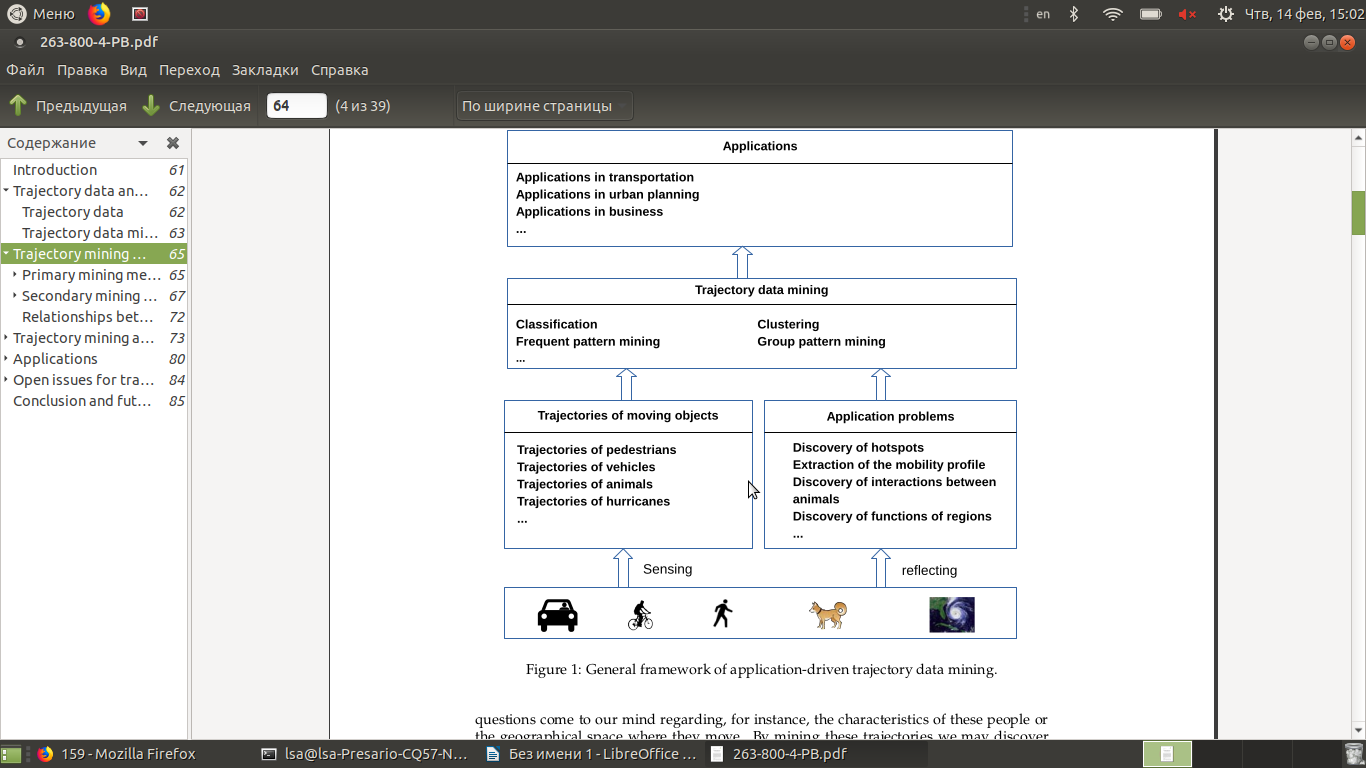
Leleko Serhiy, Palahin Volodymyr, Filipov Vitalii

**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.

**Sport track (running, cycling, walking, swimming activities)**

In process of development of the standard of living of people, physical activity and physical culture become an integral part of any person. In modern sport an important stage in training of athletes and the analysis of its results is acceptance of the effective expected solution on success of their competitive activity. The correct forecast is important for the organization of planning of training of the athlete and his adequate sports orientation.

Most methods of predicting of expert estimates of sports activity are already developed [5-7] and are applied as in amateur and in professional sport. The predicting of having an athlete result equivalent for professionals reaches up to 90%, and the success of athletes' performances fluctuates within 30-80% (on average, 50%) [5, 8]

At present, various technologies from different industries are used to improve predictions in sport [9]. To monitor the analysis of athletes, the latest developments in electronics, communications, navigation, and medicine are used. Various software tools that help control the functional systems of the body are used too.

Все аппаратные средства можно разделить на две группы:

All hardware can be divided into two groups:

* stationary (medical and sports equipment, usually installed in the laboratory for detailed testing of athletes [10-11]);
* portable (usually in the form of watches or small gadgets that are used to monitor all sports activities, and also the life cycles which are not relating to sport).

Use of stationary hardware gives a complex set of metrics [12] (physical body parameters, heart rate, lactate threshold, threshold power, maximal oxygen consumption (VO2max), anaerobic metabolism threshold (ANSP), etc.), that allows to make the analysis about a current status of the athlete, and if there is a long series of tests (for example, after each workout), it will make a precise forecast for future sporting events. But at the same time, it should be noted, that the practical use of this type of hardware is not more than two to three times a year (at the beginning of the year and before the target start). It is caused by the fact that the frequent use of special equipment, a laboratory, a testing specialist, and a high cost of money (currently, this service costs $ 100 [13]).

For current analysis, athletes (amateurs or professionals) more often use portable hardware in the form of portable pulsometers [14]. Typically, pulsometers are smart watches with built-in GPS-receivers, or cycle computers and cycling powermeters (usually for cycling) from popular manufacturers: Garmin, Polar, Suunto, Sigma [15]. In practice, these means are metric: speed (pace), distance, cadence, heart rate, heart rate zones, power (for cycling), vertical oscillation and ground contact (for running) [16-17]. Using such devices makes it easy to receive a data set of athlete's training data for further analysis and prediction of the results. It is caused by the fact that the price of such devices is relatively small (for pulsometers up to $ 500 [18]), and their use for training is quite simple.

But along with this, it is necessary to add some of the following features of their use in practice:

* The use of cycling powermeters (for the analysis of bicycle training) is possible only with a pulsometer or bike computer (smartphone), and their price is up to 1,000$ [18]. As a result for the analysis of training, generally they are used only by professional athletes, and most amateurs choose a calculated power model (software tools [19-20]).
* Some pulsemeters, just after training, can give a simple prediction based on only a few metrics: speed (pace), heart rate, time, distance, and calculated parameters (for example, VO2max).
* Using the calculated parameters leads to a strong generalization of the predict, in consequence, it will often receive a wrong data for analysis [21-23].

As for software for training analysis, their use is quite simple and only requires a smartphone existence with the built-in GPS-receiver. However, they, in most cases, use predictive parameters to predict sports results, which leads to a generalization of the received results. It is also necessary to note that some users swindle in a number of cases – to improve their result, indicate not the correct type of workout (instead of running –cycling), which is not tracked by the software (only another user in the manual mode can point to the system to the wrong type of workout).

After analyzing the arguments above, we can draw the following conclusion: for the analysis of sports training and the predict of sports results, we will use the data received from pulsometers, cycling computers, and smartphones.

On the basis of the received data (datasets) it is necessary to solve the following tasks:

* Analysis of sports training and predict of result.
* Classification of training in the form of sports (running, walking, swimming, cycling).
* Classification of training for an athlete (professional, amateur).
* Predict of the evaluated parameters (cycling power, VO2max) of the athlete in the trajectory of the movement.

**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 [24]). 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 [25, 26]. 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) [27 ]

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.

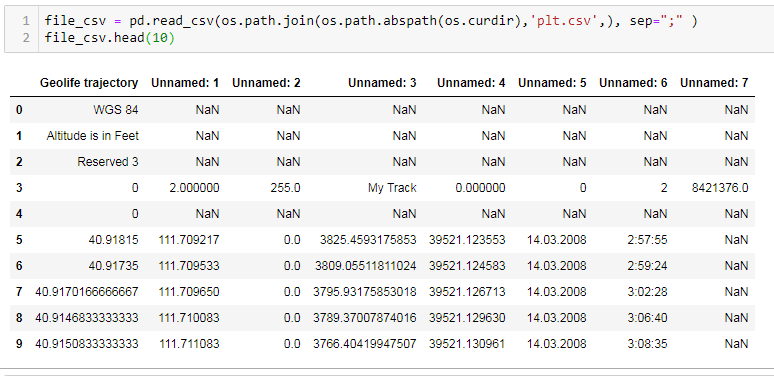
73 users have labeled their trajectories with transportation mode, such as driving, taking a bus, riding a bike and walking. There is a label file storing the transportation mode labels in each user’s folder.

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:

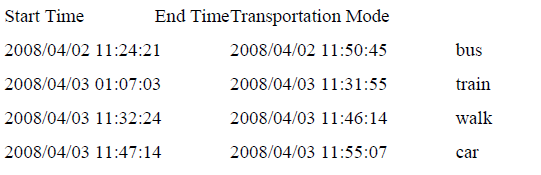
* Field 1: Latitude in decimal degrees.
* Field 2: Longitude in decimal degrees.
* Field 3: All set to 0 for this dataset.
* Field 4: Altitude in feet (-777 if not valid).
* Field 5: Date – number of days (with fractional part) that have passed since 12/30/1899.
* Field 6: Date as a String.
* Field 7: 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**

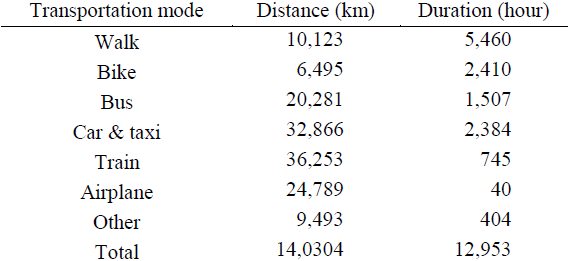
Transportation mode labels

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 [29]. 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.

**Running and Cycling Data**

Sport, track and track trackers Garmin, Polar, Suunto [15, 18] with built-in GPS navigation system (GLONASS) and pulse measurement support are used to record running, walking, cycling.

Most data that is recorded by athletes (especially amateurs) is stored in sports social networks [18-20] and is open to public access and research. In this work, we will use the data of athletes (amateurs and professionals) received from the sports social network Strava [20]. The choice of athletes and their recorded activities will be based on the following criteria:

* The athlete profile should be open and contain additional data such as temperature, pulse, and power in the activities.
* Track recordings should be made by a modern GPS device such as Garmin Fenix 3 (5), Garmin Forerunner 920 (935), Garmin Edge 520 (810, 1030) [20].
* The number of activities during the period under review should be large enough (for example, more than 100).
* The selected athlete must have different types of activities, such as cycling and running.

In general, the GPS experiments with sport data contain some parts:

* Firstly, sort data and other background information (temperature, power, wind, etc) is collected during field work.
* Secondly, the collected data is retrieved and validated (if its need)
* After retrieval, the validity and integrity of the data is checked and track logs and database are cleaned and filtered.
* During the next processing phase the valid trajectories are converted from the gpx format to csv and processed for further analysis

Figure 4 summarises the steps in the process. The processing of the data eventually leads to GPS Data Mining Processes.



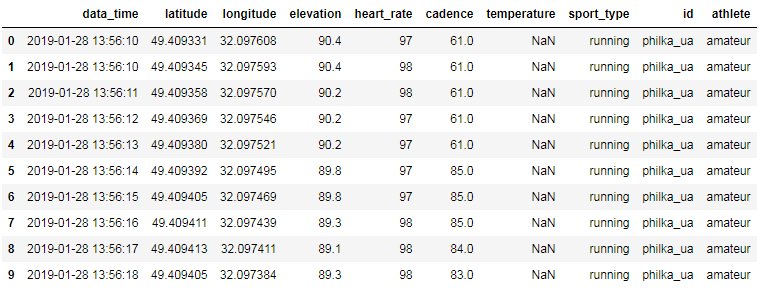
**Figure 4 – General data processing scheme for GPS experiments**

In general, GPS data (gpx files) is presented in xml format [30], which is shown on figure 5.

****

**Figure 5 – Structure of GPS data (gpx file)**

Therefore its converting in csv format (for convenient processing in Pandas DataFrame) there is an important step in preparation of GPS data. To convert GPS data into csv format, we used two libraries: gpxpy and gpx-csv-converter [31, 32]. The conversion result is shown in Figure 6.

****

**Figure 6 – Result of convertation gpx to csv files**

The basic data retrieved in csv files consist of sequentially recorded points – gps-points (usually point records, once per second) and data about the athlete:

* date – date for gps-point;
* time – time for gps-point;
* data\_time – date and time (based on first and second columns) in datatime format;
* latitude – latitude, of gps-point;
* longitude – longitude, of gps-point;
* elevation – elevation in gps-point;
* heart\_rate – heart rate in gps-point.

It is necessary to note, that in addition to the basic data on the movement of athletes (data, time, latitude, longitude, elevation), GPS data may contain additional data (depending on the GPS device):

* heart rate – heart rate in gps-point;
* cadence – frequency of steps for running, pedaling frequency for cycling;
* vertical oscillation – running metric;
* power – cycling metric;
* temperature;
* pressure.

These data are stored in the gpx file in the extensions section, under different tags (depends on the GPS device), which sometimes leads to the loss of this data in converting process. Therefore, we modified the library [31] – added tags for the main parameters that are recorded with the GPS-device.

Realization of converting gpx files to csv files in jupyter notebook [33]:

<https://github.com/PhilkaUA/gps-data-mining>

Labels: ID – name (or nick name of Athlete)

Athlete labels

* professional;
* amateur.

Sport labels: (possible kinds of sport activities)

* cycling;
* running;
* winter;
* zwift;
* other.

**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. Баландин В.И. Прогнозирование в спорте / В.И. Баландин, Ю.М. Блудов, В.А. Плахтиенко. М. : Физкультура и спорт, 1986. - 192 с.
6. Теория и методика физической культуры: программа для бакалаврского уровня высш. физкультур, образования / Матвеев Л.П., ред. – М. : Физкультура и спорт, 1997. – 76 с.
7. Фискалов В.Д. Теоретические основы и организация подготовки спортсменов: учебное пособие / Фискалов В.Д. – В.: ФГОУВПО «ВГАФК». – 2010. – 278 с. – Режим доступа: <https://goo.gl/oSDAow>
8. Aidman E.V. Role of Personality Attributes in Long-term Predictions of Achievement in Sport: The Journal of the American Board of Sport Psychology, 2007, – volume 1. – Access mode: <https://www.researchgate.net/publication/284098744_Attribute-based_selection_for_success_The_role_of_personality_attributes_in_long_term_predictions_of_achievement_in_sport>
9. Сорокин В.П. Цифровые технологии в практике физической подготовки в высшем военно-учебном заведении / В.П. Сорокин, А.А. Михайлович, И.Д. Евгеньевич, И.Ю. Тимофеев, И.М. Колбая: Ученые записки университета им. П.Ф.Лесгафта. – 2018. – №7(161). – с. 258-262. – Режим доступа: <https://goo.gl/nfKuEE>
10. Felix Lebed. Complex Sport Analytics (Routledge Research in Sport and Exercise Science) 1st Edition /Routledge Research in Sport and Exercise Science. – Routledge; 1 edition. – 2017, May 19. – 318 p. – Access mode: <https://goo.gl/szC9Yg>
11. Шевко Н.Б. Комплексное тестирование функциональной готовности спортсменов / Н. Б. Шевко // Проблемы здоровья и экологии. – 2007. – № 3 (13). – С. 116-122. – Access mode: <http://elib.gsmu.by:80/xmlui/handle/GomSMU/1315>
12. Run and travel: [сайт]. – Режим доступа: <http://run-and-travel.com/functionalnoe-testirovanie-sportsmenov-lubiteley/>
13. VO2 maxlab: [сайт]. – Режим доступа: <http://vo2maxlab.in.ua/services/>
14. Live Science – Health: [сайт]. – Режим доступа: <https://www.livescience.com/42220-heart-rate-monitors.html>
15. Для чего нужен пульсометр в спорте: [сайт]. – Режим доступа: <https://www.iphones.ru/iNotes/pulsometer-benefits>
16. Luc Léger. Heart Rate Monitors: Validity, Stability, and Functionality / Luc Léger, Madeleine Thivierge. The Physician and Sportsmedicine. – Volume 16, Issue 5. – 2016, 12 Jul. – p. 143-151.

Access mode: <https://doi.org/10.1080/00913847.1988.11709511>

1. Running metrics: [сайт]. – Режим доступа: <https://www.wareable.com/running/your-running-watch-explained-what-does-cadence-vertical-oscillation-and-hr-max-mean>
2. Cycing and running store: [сайт]. – Режим доступа: <https://www.bike-discount.de/en/bike-electronics>
3. Trainning Peaks – Losing and Regaining Fitness: [сайт]. – Режим доступа: <https://www.trainingpeaks.com/blog/losing-and-regaining-fitness/>
4. Sport social network: [сайт]. – Режим доступа: <https://www.strava.com>
5. Forum Garmin: [сайт]. – Режим доступа: <https://forums.garmin.com/forum/on-the-trail/wrist-worn/fenix-2/62948->
6. Forum Garmin: [сайт]. – Режим доступа: <https://forums.garmin.com/forum/on-the-trail/wrist-worn/fenix-2/60823-vo2-and-race-predictor-accuracy?84743-Vo2-and-race-predictor-accuracy>=
7. Forum Garmin: [сайт]. – Режим доступа: <https://forums.garmin.com/forum/into-sports/running/forerunner-620-aa/65528-?91743-What-are-your-FR620-VO2-max-estimates-race-prediction-times>=
8. Kahtan Aziz. Smart real-time healthcare monitoring and tracking system using GSM/GPS technologies / Kahtan Aziz, Saed Tarapiah, Salah Haj Ismail, Shadi Atalla //Materails of 3rd MEC International Conference on Big Data and Smart City (ICBDSC). – IEEE Xplore: 28 April 2016. – ISBN Information: INSPEC Accession Number: 15955334.

Access mode: <https://ieeexplore.ieee.org/abstract/document/7460394/authors#authors>

1. 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: Building Social Networks Using Human Location History: [сайт]. – Режим доступа: <https://goo.gl/F8EUsf>
4. GeoLife GPS Trajectories: [сайт]. – Режим доступа: <https://goo.gl/HUy1iS>
5. GPS Files: [сайт]. – Режим доступа: <https://www.topografix.com/GPX/1/1/>
6. Python libraries [сайт]. – Режим доступа: <https://pypi.org/project/gpxpy/>
7. Python libraries [сайт]. – Режим доступа: <https://pypi.org/project/gpx-csv-converter/>
8. Jupyter nootebook – converter gpx-csv – [сайт].

Режим доступа: <https://goo.gl/B2aap4>

**TRAJECTORY DATA MINING:**

**Researching trajectory data by GPS tracks**

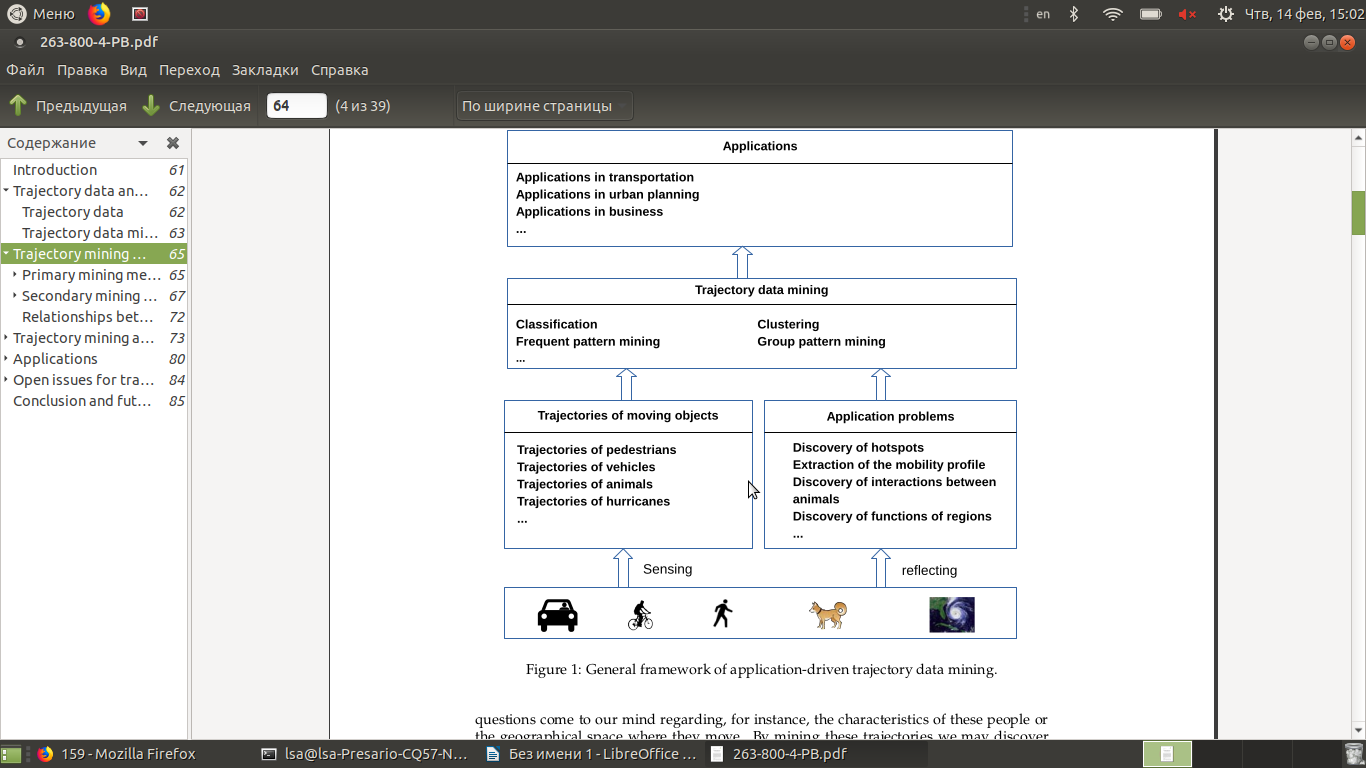
Leleko Serhiy, Palahin Volodymyr, Filipov Vitalii

**1. Define the problem**

Развитие информационных и коммуникационных технологий, беспроводной связи, дает нам возможность получить данные, которые содержат изменяющиеся во времени географические координаты. Их анализ может привести к решению важных исследовательских задач в различных областях, таких как городское планирование, транспорт, поведенческая экология, анализ спортивных тренировок, наблюдение и безопасность.

**Vehicles tracking (машины, такси, грузовые перевозки)**

В настоящее время много гражданских машин, коммерческого и муниципального транспорта (автобусы, такси), оборудуют GPS-приемниками для отслеживания их траектории движения [1]. На основе полученных данных зачастую решают следующие задачи: оптимизацию траффика, логистику перемещения транспорта (для коммерческих перевозок), анализ траектории движения и отклонения от нее (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.

**Sport tracking (running, cycling, walking, swimming activities)**

По мере развития уровня жизни людей, физическая активность и физическая культура становится неотъемлемой частью любого человека. В современном спорте важным этапом в подготовке спортсменов и анализе его результатов является принятие эффективного прогнозного решения об успешности их соревновательной деятельности. Правильный прогноз важен для организации планирования подготовки спортсмена и его адекватной спортивной ориентации.

Большинство методов прогнозирования экспертных оценок спортивной активности уже разработаны [5-7] и применяются как в любительском так и профессиональном спорте. Прогноз о наявности у спортсмена результатов эквивалентных для профессионалов достигает до 90%, а точность прогнозов успешности выступлений спортсменов колеблется в пределах 30-80% (в среднем 50%) [5, 8].

В нынешнее время для улучшения прогнозов в спорте используют различные технологии из различных отраслей [9]. Для мониторинга анализа действий спортсменов используют последние разработки в области электроники, связи, навигации и медицины. Так же используются различные программные средства, которые помогают контролировать функциональные системами организма.

Все аппаратные средства можно разделить на две группы:

* стационарные (медицинское, спортивное оборудование, как правило установленное в лаборатории для проведения детального тестирования спортсменов [10-11]);
* портативные (как правило в виде часов или небольших гаджетов которые используются для мониторинга всех спортивных активностей, а также жизненных циклов, не относящихся к спорту).

Использование стационарных аппаратных средств дает комплексный набор метрик [12] (physical body parameters, heart rate, lactate threshold, threshold power, maximal oxygen consumption (VO2max), anaerobic metabolism threshold (ANSP), etc), что позволяет сделать анализ о текущем состоянии спортсмена, а при наличии длинной серии тестов (например, после каждой тренировки), позволит сделать точный прогноз для будущих спортивных событий. Но вместе с этим, следует заметить, что использование на практике такого типа аппаратных средств, бывает не больше двух-трех раз за год (в начале года и перед целевым стартом). Это обусловлено тем, что требуется частое использование специальное оборудования, лаборатории, специалиста по тестированию и больших затрать по деньгам (на данный момент эта услуга стоит 100$ [13]).

Для поточного анализа спортсменами (любителями или профессионалами) чаще используются портативные аппаратные средства в виде портативных пульсометров [14]. Как правило пульсометры – это часы (smart watches) с встроенным gps-приемником, или велокомпьютеры и cycling powermeters (обычно для велоспорта) от популярных производителей: Garmin, Polar, Suunto, Sigma [15]. На практике с помощью этих средств получают такие метрики: speed (pace), distance, cadence, heart rate, heart rate zones, power (for cycling), vertical oscillation and ground contact (for running) [16-17]. Использование таких устройств, позволяет достаточно просто получить набор данных тренировок спортсмена, для подальшего анализа и прогноза результатов. Это обусловлено тем, что цена таких устройств относительно не велика (для пульсометров до 500$ [18]), а их использование для тренировок достаточно просто. Но вместе с этим, следует добавить несколько следующих особенностей их использования на практике:

* Использование cycling powermeters (для анализа велосипедних тренировок) возможно только совместно с пульсометром или велокомпьютером (смартфоном), а их цена составляет до 1000$ [18]. Как следствие для анализа тренировок, восновном их используют только профессиональные спортсмены, а большинство любителей выбирают расчетную модель мощности (программные средства [19-20]).
* Некоторые пульсометры, сразу после тренировки могут давать простой прогноз, основываясь только на нескольких метриках: speed (pace), heart rate, time, distance и расчетных параметрах (например VO2max).
* Использование расчетных параметров приводит к сильному обобщению прогноза, в следствии чего, часто получают ошибочные данные для анализа [21-23].

Что касается программных средств для анализа тренировок, то их использование достаточно просто и требует только наличие смартфона со встроенным gps-приемником. Однако они, в большинстве случаев, для прогноза спортивных результатов используют расчетные параметры, что приводит к обобщению полученных результатов. Также нужно отметить, что некоторые пользователи в ряде случаев жульничают – для улучшения своего результата, указывают не верный тип тренировки (вместо бега – велосипед), что никак не отслеживается программным средством (только другой пользователь в ручном режиме может указать системе на неверный тип тренировки).

Проанализировав выше приведенные доводы, можем сделать следующий вывод: для анализа спортивных тренировок и прогноза спортивных результатов, будем использовать данные полученные от пульсометров, велокомпьютеров, смартфонов.

На основе полученных данных (datasets) необходимо решить следующие задачи:

* Анализ спортивной тренировки и прогноз результата.
* Классификация тренировки по виду спорта (running, walking, swimming, cycling)
* Классификация тренировки по спортсмену (профессионал, любитель)
* Прогноз расчетных показателей (cycling power, VO2max) спортсмена по траектории движения.

**2. Data acquisition**

Данные о траекториях движения или перемещения исследуемых объектов обычно получают с помощью устройств оборудованных GPS (или GLONASS) модулями (реже GSM/WiFi модулями [24]). В наших исследованиях в качестве источника gps-данных, мы выберем те навигационные устройства, работа которых, базируется на основе GPS (GLONASS) технологий, поскольку их точность достаточно высока – до 10 м на открытой местности, для гражданских устройств [25, 26]. А данные полученные от GSM/WiFi модулей рассматривать не будем, в связи с их низкой точностью, как для навигационных устройств.

**Machines and Vehicles Data**

Для исследования траекторий движения **транспорта** (машин, автобусов, такси, пешеходов) мы воспользуемся GPS trajectory dataset that was collected in (Microsoft Research Asia) Geolife project by 182 users in a period of over five years (from April 2007 to August 2012) [27].

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.

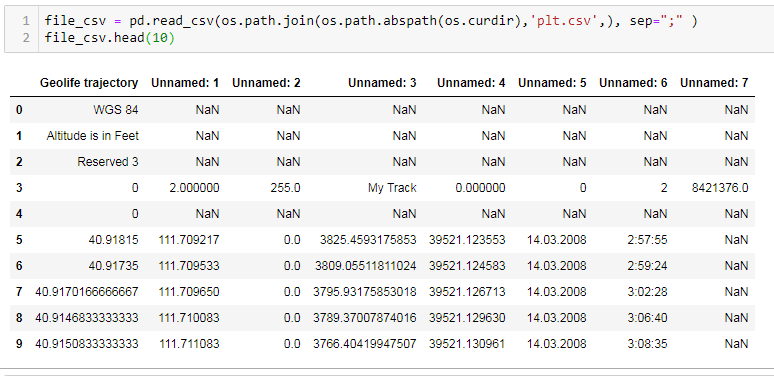
73 users have labeled their trajectories with transportation mode, such as driving, taking a bus, riding a bike and walking. There is a label file storing the transportation mode labels in each user’s folder.

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:

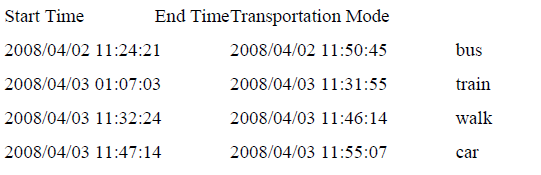
* Field 1: Latitude in decimal degrees.
* Field 2: Longitude in decimal degrees.
* Field 3: All set to 0 for this dataset.
* Field 4: Altitude in feet (-777 if not valid).
* Field 5: Date – number of days (with fractional part) that have passed since 12/30/1899.
* Field 6: Date as a String.
* Field 7: 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**

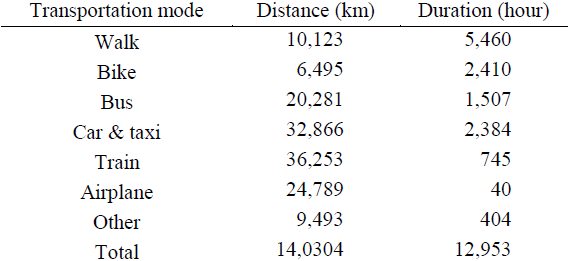
Transportation mode labels

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 [29]. 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.

**Running and Cycling Data**

Для записи траекторий движения **в спорте** (running, walking, cycking) используются спортивные трекеры Garmin, Polar, Suunto [15, 18] с встроенной навигационной системой GPS (GLONASS) и поддержкой измерения пульса.

Большинство данных, что записывается спортсменами (особенно любителями) сохраняется в спортивные социальные сети [18-20] и открыты для общего доступа и исследований. В этой работы мы будем использовать данные спортсменов (любителей и профессионалов) полученные со спортивной социальной сети Strava [20]. Выбор спортсменов и их записанных активностей будем производить опираясь на следующие критерии:

* Профиль спортсмена должен быть открытым и содержать в активностях дополнительных данных, например, температуры, пульса, мощности.
* Запись трека должна производиться современным GPS-устройсвом, например, Garmin Fenix 3 (5), Garmin Forerunner 920 (935), Garmin Edge 520 (810, 1030).
* Количество активностей за исследуемый период, должно быть достаточно большим (например, больше чем 100).
* Выбранный спортсмен должен иметь разные типы активностей, например cycling and running.

In general, the GPS experiments with sport data contain some parts:

* Firstly, sort data and other background information (temperature, power, wind, etc) is collected during field work.
* Secondly, the collected data is retrieved and validated (if its need)
* After retrieval, the validity and integrity of the data is checked and track logs and database are cleaned and filtered.
* During the next processing phase the valid trajectories are converted from the .gpx format to .csv and processed for further analysis

Figure 4 summarises the steps in the process. The processing of the data eventually leads to GPS Data Mining Processes.



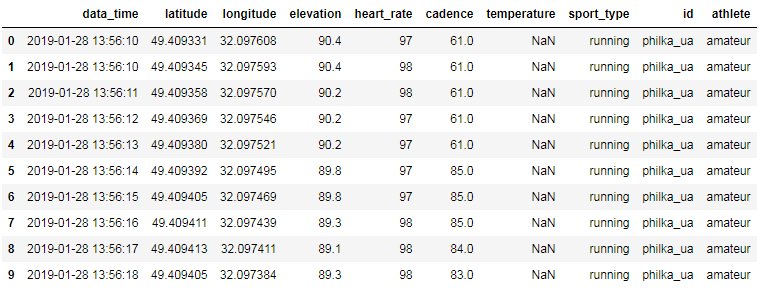
**Figure 4 – General data processing scheme for GPS experiments**

В общем виде GPS data (gpx files) представлены в xml формате [30], что показано на figure 5.

****

**Figure 5 – Structure of GPS data (gpx file)**

Поэтому важным этапом в подготовке GPS data, есть ее конвертация в csv format (для удобной обработки в Pandas DataFrame). Для конвертирования GPS data в csv format мы использовали две библиотеки: gpxpy и gpx-csv-converter [31, 32]. Результат конвертирования показан на figure 6.

****

**Figure 6 – Result of convertation gpx to csv files**

Основные данные полученные в **csv files** состоят из последовательно записанных точек – gps-points (обычно запись точек, ведётся раз в секунду) и данных о спортсмене:

* date – date for gps-point;
* time – time for gps-point;
* data\_time – date and time (based on first and second columns) in datatime format;
* latitude – latitude, of gps-point;
* longitude – longitude, of gps-point;
* elevation – elevation in gps-point;
* heart\_rate – heart rate in gps-point.

Необходимо выделить, что кроме основных данных о перемещении спортсменов (data, time, latitude, longitude, elevation), в зависимости от GPS-device, GPS data могут содержать дополнительные данные:

* heart rate – heart rate in gps-point;
* cadence – frequency of steps for running, pedaling frequency for cycling;
* vertical oscillation – running metric;
* power – cycling metric;
* temperature;
* pressure.

Эти данные хранятся в gpx file в разделе extensions, под разными тегами (зависит от GPS-device), что иногда привод к тому, что при конвертации эти данные теряются. Поэтому мы модифицировали библиотеку [31] – добавили теги для основных параметров которые записываются с GPS-device.

Реализация конвертации gpx files to csv files in jupyter notebook [33]:

<https://github.com/PhilkaUA/gps-data-mining>

Labels: ID – name (or nick name of Athlete)

Athlete labels

* professional;
* amateur.

Sport labels: (possible kinds of sport activities)

* cycling;
* running;
* winter;
* zwift;
* other.