

Data-driven approach for anomaly detection of real GPS trajectory data

Emir Barucija

*Faculty of Electrical Engineering
University of Sarajevo
Sarajevo, Bosnia and Herzegovina
ebarucija1@etf.unsa.ba*

Amra Mujcinovic

*Faculty of Electrical Engineering
University of Sarajevo
Sarajevo, Bosnia and Herzegovina
amujcinovi3@etf.unsa.ba*

Berina Muhovic

*Faculty of Electrical Engineering
University of Sarajevo
Sarajevo, Bosnia and Herzegovina
bmuhovic1@etf.unsa.ba*

Emir Zunic

*Faculty of Electrical Engineering
University of Sarajevo
Sarajevo, Bosnia and Herzegovina
emir.zunic@etf.unsa.ba*

Dzenana Donko

*Faculty of Electrical Engineering
University of Sarajevo
Sarajevo, Bosnia and Herzegovina
ddonko@etf.unsa.ba*

Abstract—The last decade was marked by rapid growth and development of technology. One example of that is the automotive industry. This industry has made an enormous progress, and its main goal is to achieve safer and better driving. The vehicle incorporates GPS devices that send information about the current location and speed of the vehicle. Large amounts of collected data can be used in companies for tracking vehicles and various analysis and statistics. Sometimes, however, GPS data is not accurate. In this paper, the potential of real data sets will be used to analyze possible anomalies that may occur when reading GPS position of vehicles. The approach for solving this problem used in this paper consists of calculating distance and time, based on GPS measurements, then calculating average speed based on these two values, and comparing that speed with the speed given by GPS device.

Index Terms—GPS; anomaly; speed; distance; time

I. INTRODUCTION

GPS is one of the most useful inventions of modern technology. It is built into every mobile phone and it is a standard addition to many vehicles. Even at the very beginning of the automotive industry, it was a great challenge to move from point X and reach point Y without any problems [1]. Originally, the problem with getting around in space was solved by using the printed maps until the GPS came out. GPS is a modern solution that makes it easier to navigate and reach the desired destinations, while saving time.

In most cases, it is possible to rely on the GPS to get the required position. However, this is not 100% correct, so reading the current position might be unavailable. There are potential cases where GPS position reading fails, which will be presented below.

One of the possible problems are various environmental conditions. As an example, sunlight influence can be taken into account, which can cause distortion of the signal, although this happens sporadically. In addition, there is also the influence of human intervention. Cheap GPS devices were built into cars,

which are not precise enough and do not have enough quality. Design of the screen that displays the current position and tracks path on the map can also affect GPS readings. Antenna position in the vehicle is one of the factors that contributes to proper functioning of the GPS system. Poor mobile network coverage is another cause of anomaly when reading GPS location. The consequence of this is the inability to send current position due to a weak or non-existent mobile signal. Generally, location readings occur through mobile network, so if there is no signal, anomalies are inevitable. It is well known that GPS reading consumes a lot of battery power, so poor battery status can lead to inability to gain the current GPS position. More about these problems can be found in [2].

Because of all aforementioned problems, anomalies when reading the GPS position may occur. The consequences of anomalies are unreliability and imprecision, which can cause many other problems. In order to reduce or eliminate the number of anomalies, it is necessary to analyze how often they occur. The aim of this paper is to analyze the appropriate data sets and use tools needed to get the information about percentage of anomalies in the data set.

Next section is about related work in this field. Most of the papers mentioned in next section are focused on anomaly detection. Some of them only detect GPS issues, but some solve problems using proposed algorithms and approaches, which are usually based on cleaning data. The third section is Case Study. The purpose of that section is the demonstration of proposed approach for anomaly detection. The section shows basic steps important for detecting issues with GPS tracking on a real data set. Section after that, labeled Results, is closely related to Case Study section. It demonstrates the results obtained using the proposed approach. Results are presented using images, tables and diagrams for better understanding. Last section contains conclusion, as well as open opportunities for further refinement.

II. RELATED WORK

Since GPS devices are used in practice a lot, the problem of anomaly detection can be quite significant to some companies which track their vehicles using GPS devices. Because of that, there are lots of papers dealing with this topic. In this section, four papers will be mentioned, along with approaches and methods used for detection. Some of those papers also proposed a way of correcting GPS data, which will also be described in short.

A. Anomaly Detection in GPS Data Based on Visual Analytics

Modern machine learning provides numerous approaches for modeling data and extracting critical information. In the paper [3], researchers have combined two types of anomaly detection, using visualization and human-computer interaction. This paper presents GPS Visual Analytics System, a system that detects anomalies in GPS data by using visual analytics. The CRF (Conditional Random Field) model is used as a component for machine learning to detect anomaly when monitoring GPS readings. The essence of the work is focus on data analysis and detection of GPS urban taxi anomalies.

B. GeoSClean: Secure Cleaning of GPS Trajectory Data using Anomaly Detection

GPS and location based services are one of the most used technologies today. Various applications, like Uber and Maps, use them. GPS is present in many devices, such as smartphones and IoT (Internet of Things). It is a fact that such systems on the cloud can be accessed almost uninterruptedly. On the other hand, there are disadvantages, which are, among other things, security and privacy, due to cloud attacks. That is why many users do not want to share their location. Because of these reasons, the GeoSClean method is used in the paper [4]. By using this method, GPS trajectory can be cleaned from anomalies, but it can also confidentially keep user's location. There are many ways in which anomalies can be detected, and this paper performs analysis of speed, time and acceleration.

C. Time Series Data Cleaning: From Anomaly Detection to Anomaly Repairing

When working with GPS data and sensor readings, many errors may occur. It is very important to discover the existence of these anomalies. There are methods that only detect anomalies, but they do not solve them. It is possible to detect errors in the data, but the problem of unreliability is still present. The approach in the paper [5] is based on the simultaneous detection and repair of errors, which includes the following steps: (1) a novel framework of iterative minimum repairing (IMR) over time series data, (2) explicit analysis on convergence of the proposed iterative minimum repairing, and (3) efficient estimation of parameters in each iteration. When using the iterative account over the real data set, with n data points, the complexity of the algorithm changes from $O(n)$ to $O(1)$. Researchers have shown that iterative computation proved to be successful.

D. Road Traffic Anomaly Detection via Collaborative Path Inference from GPS Snippets

Detecting anomalies when analyzing GPS snippets is critical in an urban computer science due to background events. GPS data has brought numerous problems, in terms of detecting anomalies, that are very challenging to solve. For solving the problem the following two-step solution is used: a Collaborative Path Inference (CPI) model and a Road Anomaly Test (RAT) model [6]. CPI model performs path inference incorporating both static and dynamic features into a Conditional Random Field (CRF). RAT calculates degree of anomalies for each segment of the road at certain time intervals. Paper [6] presents a real data set that includes driving data of eight thousand taxi cars within a month. The results show the advantages the method described in [6] beyond other techniques.

III. CASE STUDY

Settings of GPS device that produced the data set used in this paper are shown in the Fig. 1. As the figure states, when the vehicle is not moving, it needs to send GPS measurement at least every 3600 seconds. If the vehicle is moving, it needs to send measurement after the distance travelled is at least 350 meters, or after the angle of the wheel changes by at least 14 degrees, whichever condition is met first. If none of that happens under 33 seconds, then the device shall send measurement after those 33 seconds (for example, if the traffic light is red etc).

The data set (available upon querying authors on e-mail) has 6853 rows, one for every GPS measurement, and following attributes: vehicle number, timestamp, latitude (in degrees), longitude (in degrees), vehicle speed and altitude. First step is to add new attributes (columns) to the data set, which are: total time in seconds (in Unix/POSIX time) and latitude and longitude in radians are created.

The block diagram of the approach used in this paper is shown in Fig. 2.

Idea of the approach is to take every two successive rows from the data set, which represent two successive GPS measurements, and to calculate distance and time between them. Time is simply calculated as the difference of total time for second and first measurement. Regarding the distance, it is first needed to do conversion from Polar to Cartesian coordinates, which is given by equations (1) - (3):

$$x = r * \cos(\text{latitude}) * \sin(\text{longitude}) \quad (1)$$

$$y = r * \sin(\text{latitude}) \quad (2)$$

$$z = r * \cos(\text{latitude}) * \cos(\text{longitude}) \quad (3)$$

Here, r is the total distance from the center of the Earth. This is obtained as the sum of the average Earth radius [7], and the altitude of the GPS measurement. After that, the distance is calculated by Euclidean formula, given in the equation (4):

$$\text{distance} = \sqrt{(\Delta x)^2 + (\Delta y)^2 + (\Delta z)^2} \quad (4)$$

| Home | | |
|--|-------------|------|
| Vehicle on Stop | | |
| Min. period: | 3600 | sec. |
| | | |
| Min. saved records : | 1 | |
| Send period: | 30 | sec. |
| Vehicle Moving | | |
| Min. period: | 33 | sec. |
| Min. distance: | 350 | m. |
| Min. angle: | 14 | deg. |
| Min. saved records : | 10 | |
| Send period: | 30 | sec. |
| GPRS Context Week Time | | |
| Week days | Time of day | |
| <input checked="" type="checkbox"/> M | 00:00 | |
| <input checked="" type="checkbox"/> T | 00:10 | |
| <input checked="" type="checkbox"/> W | 00:20 | |
| <input checked="" type="checkbox"/> T | | |
| <input checked="" type="checkbox"/> F | | |
| <input checked="" type="checkbox"/> Sa | | |
| <input checked="" type="checkbox"/> Su | | |
| Check All Clear All | | |
| Week days | Time of day | |
| <input checked="" type="checkbox"/> M | 00:00 | |
| <input checked="" type="checkbox"/> T | 00:10 | |
| <input checked="" type="checkbox"/> W | 00:20 | |
| <input checked="" type="checkbox"/> T | | |
| <input checked="" type="checkbox"/> F | | |
| <input checked="" type="checkbox"/> Sa | | |
| <input checked="" type="checkbox"/> Su | | |
| Check All Clear All | | |

Fig. 1: Settings of GPS device.

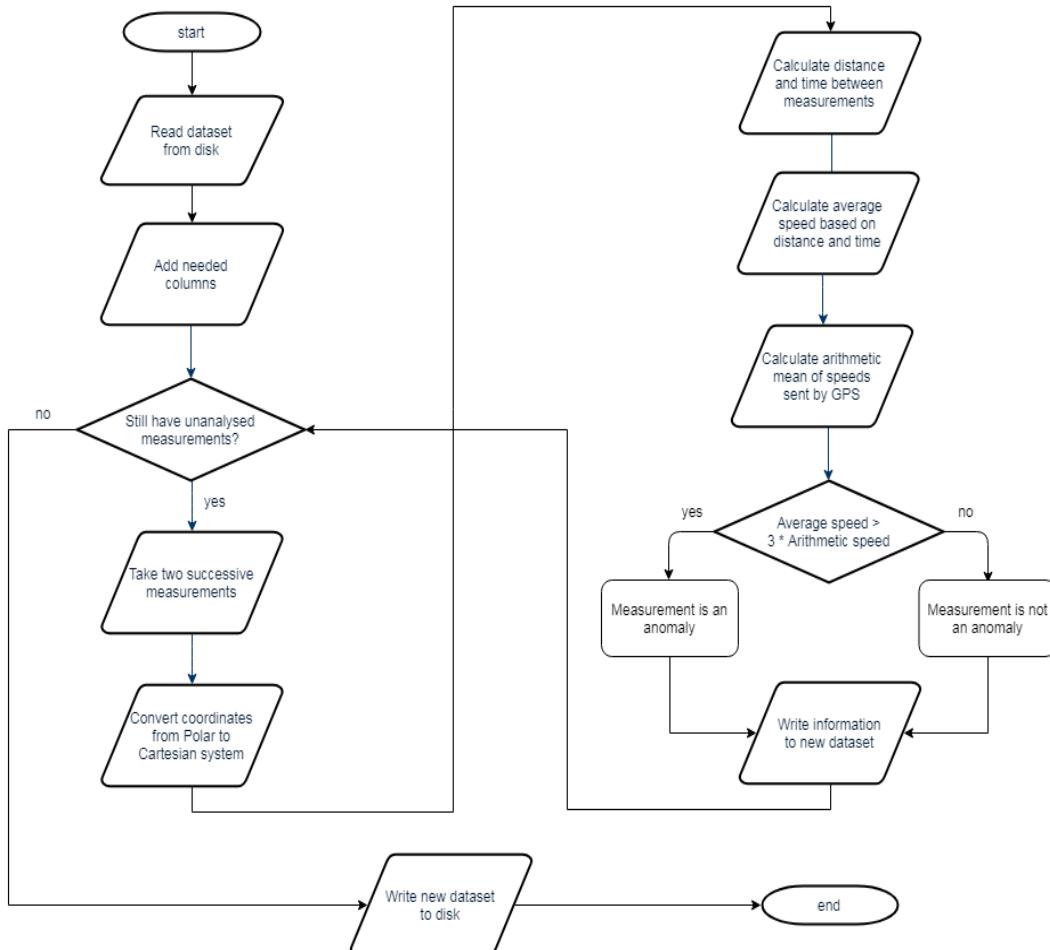


Fig. 2: Block diagram of the approach.

Δx , Δy and Δz are differences in x, y and z coordinates of successive GPS measurements.

Next step is to calculate average speed, which is done by a well known physics formula, given in equation (5):

$$v_{\text{average}} = \text{distance}/\text{time} \quad (5)$$

This speed represents the average speed of the vehicle between these two measurements. If the average speed calculated this way differs too much from the arithmetic mean of speeds sent by GPS, then, these measurements can be considered as anomalies. The reason for this is that, for these measurements, the distance travelled is too long for such a short interval of time (for example: the vehicle travels 150 meters in 1 second, which results in a speed of 150 m/s or 540 km/h, and that does not make any sense, because vehicles that have GPS devices built in can not achieve such speeds).

In analysis of ECG signals, upper and lower frequency thresholds are defined, where upper frequency is 3 times greater than lower. Everything which is greater than the upper threshold should be removed [8]. Following standard processes of removing high frequencies, on a similar principle, the factor giving the quotient of average speed and arithmetic mean of speeds from GPS was defined. If this factor is greater than 3, these measurements are claimed to be anomalies. Due to some imperfections of GPS measurements, if the value 2 was used for this factor, instead of 3, some regular GPS measurements could be classified as anomalies. That is the reason why the approach uses value 3 for this factor, for distinction between anomalies and non-anomalies.

For obtaining cases where GPS device did not send measurement and it was supposed to, the following procedure for detection is used. If the distance is more than 350 meters, or time between two measurements is more than 33 seconds, then definitely at least one measurement is missing. These cases are detected and reported as missing measurements.

IV. RESULTS

The data set used for this paper contains 6853 rows, which represent measurements from GPS device. The main goals are to figure out which of these measurements are anomalies (GPS sent wrong coordinates), and which measurements are eventually missing. The GPS device used is working properly, but sometimes, due to the conditions mentioned above, it produces anomalies. Only one GPS device is used, because all GPS devices generally work the same way, by recording basic parameters (geographical coordinates, time etc) which are used in this paper as well as others.

New data set is made from the original one, and it has 6852 rows, one less than the original one, because every two rows from the original data set make one row in the new data set. The approach described above was implemented in R language, and results regarding percentage of anomalies are presented in Fig. 3.

As it can be seen in the Fig. 3, the approach detected 140 anomalies, out of 6852 rows in the data set, which makes about 2% of the total data. Some of those anomalies, with

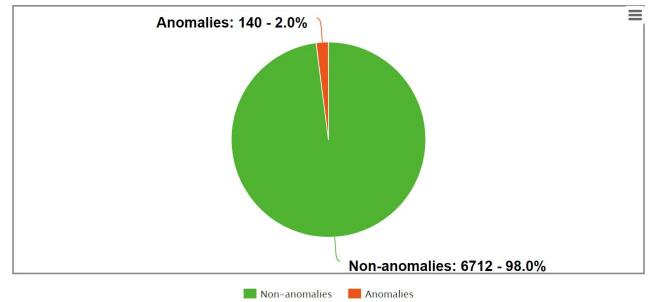


Fig. 3: Percent of anomalies in the data set.

their specific values of distances, times and speeds, can be seen in the Table I.

TABLE I: Anomalies in the data set.

| Distance | Time | Average speed m/s | Average speed km/h | Arithmetic mean speed | Speed GPS1 | Speed GPS2 |
|----------|------|-------------------|--------------------|-----------------------|------------|------------|
| 115.8 | 1.13 | 102.4 | 368.9 | 106.0 | 104 | 108 |
| 522 | 1.1 | 461.9 | 1663.0 | 59.5 | 119 | 0 |
| 648.4 | 11.1 | 58.2 | 209.7 | 56.0 | 0 | 112 |
| 29.8 | 0.1 | 211.9 | 763.1 | 101.5 | 102 | 101 |
| 22.3 | 1.1 | 19.8 | 71.3 | 22.5 | 21 | 24 |
| 8.3 | 1.1 | 7.3 | 26.5 | 7 | 7 | 7 |
| 5.0 | 1.0 | 4.7 | 17.0 | 4.5 | 0 | 9 |
| 7.0 | 1.1 | 6.2 | 22.4 | 7 | 7 | 7 |
| 16.6 | 2.0 | 8.0 | 28.9 | 8.5 | 7 | 10 |
| 20.1 | 1.1 | 17.8 | 64.3 | 18.5 | 14 | 23 |

In the first row of Table I, the distance is about 116 meters, travelled in 1.13 seconds, which makes an average speed of 369 km/h. The arithmetic mean of speeds sent by GPS device is 106 km/h. This case satisfies the condition that the average speed is more than 3 times greater than arithmetic mean of speeds, so this measurement is classified as an anomaly.

Average speed in the second row is 1663 km/h (does not make sense at all), and arithmetic mean of speeds is about 60 km/h, so this is even better example of an anomaly.

Figure 4. shows the graph containing average speed and arithmetic mean speed, where average speed is represented as a blue line, and arithmetic mean speed as a red line. The graph presents 200 measurements taken from the data set, where x axis is a measurement number, and y axis shows value of speed in km/h. Anomalies are presented as circles, where blue circle refers to the average speed calculated by the approach, and red circle to arithmetic mean of speeds measured by GPS.

Result of finding missing measurements in data set is presented on Fig. 5. As it can be seen on Fig. 5, the approach detected 62 cases where GPS measurements are missing. This could be due to the GPS not working properly, GPS has taken but has not sent the measurement, measurement has been sent, but it has not been received on the other end, or measurement

is not stored properly in the database. Some of those cases of missing measurements, with specific values of distances, times and speeds, can be seen in the Table II.

In Table II, in all rows, the distance is greater than 350 meters, and, in addition, in some rows, the time is greater than 33 seconds, so all these rows represent cases where GPS measurements are missing.

TABLE II: Missing measurements in the data set.

| Distance | Time | Average speed m/s | Average speed km/h | Arithmetic mean speed | Speed GPS1 | Speed GPS2 |
|----------|------|-------------------|--------------------|-----------------------|------------|------------|
| 522.0 | 1.1 | 461.9 | 1663.0 | 59.5 | 119 | 0 |
| 648.4 | 11.1 | 58.2 | 209.7 | 56.0 | 0 | 112 |
| 545.0 | 33.8 | 16.1 | 58.0 | 3.5 | 7 | 0 |
| 584.6 | 33.0 | 17.6 | 63.7 | 3.5 | 7 | 0 |
| 1091.9 | 13.1 | 83.1 | 299.3 | 1.0 | 0 | 2 |
| 567.1 | 33.5 | 16.8 | 60.8 | 7.5 | 15 | 0 |
| 418.6 | 9.2 | 45.4 | 163.6 | 7.0 | 0 | 14 |
| 10555.6 | 2.2 | 4784.9 | 17225.9 | 55.0 | 52 | 58 |
| 10534.6 | 5.1 | 2055.9 | 7401.4 | 52.5 | 58 | 47 |
| 584.0 | 33.9 | 17.2 | 62.0 | 3 | 6 | 0 |

Also, the approach was made to give graphical representation of anomalies on a real map, so one of these anomalies can be seen on Fig. 6. There are many regular measurements on the figure, represented by grey circles, and one measurement (red circle) that obviously deviates from others, which represents an anomaly.

The approach achieved its goal, succeeding at detecting both anomalies and missing measurements.

When comparing results of this approach with ones described in [9], it can be seen that the results obtained are close to each other. In [9], total number of anomalies was 0.46%. The approach presented in this paper used the same data set as [9], and total number of anomalies was 0.5%. As it can be

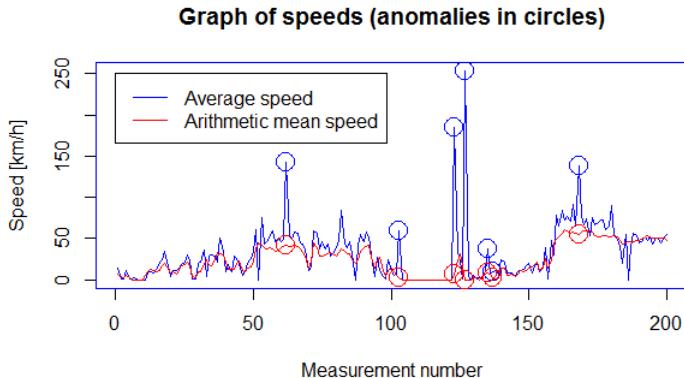


Fig. 4: Graph containing both speeds, where anomalies are represented as circles.

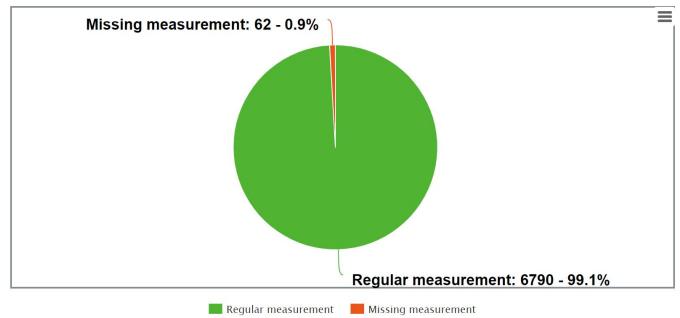


Fig. 5: Percent of missing measurements in the data set.

seen, these two approaches give similar results over the same data set.

Correctness of GPS data can be very important for various functionalities. One of them is to properly determine needed real parameters of Vehicle Routing Problem (VRP) [10].

This approach is created in the way that it can quite easily be combined with some machine learning and artificial intelligence [11], so that it first has a stage of learning whether the measurement is an anomaly, based on the actual data, and after the learning phase, it can be tested to see whether it should predict things successfully.

V. CONCLUSIONS AND FUTURE WORK

In the first section of this paper, an introduction to this topic was presented, explaining the problem and its history. After that, there was a list of papers with some previous work on this topic, and things they used to solve the problem. Then, in the Case Study section, the approach used for detection was explained, which, in short, consisted of finding distance travelled by a vehicle in a specific amount of time, then calculating average speed and comparing that speed with speeds sent by GPS device. By comparing those two speeds, approach decided whether the measurement was an anomaly or not. Also, considering the missing measurements, specifications from GPS device stated conditions on when the measurement should be sent, so approach found cases where conditions were met, but the measurement was missing.

After that, the results were presented, which consisted of statistics and graphs about anomalies and missing measurements, along with pictures of anomalies on real maps, and tables showing exact values of distance, time, average speed and arithmetic mean of speeds from GPS device, for anomalies and missing measurements.

As a conclusion, it can be said that the goal is met, an approach is proven to be successful, it can detect anomalies in GPS data, and also detect cases when the GPS measurement was supposed to be sent, but, it was not sent.

From a practical point of view, it might be even more interesting to use anomalies detection to improve existing inexpensive GPS devices for greater reliability.

In the future work on this topic, it would be good to try this approach on other data sets, possibly with some that have the information about whether measurement are anomalies indeed

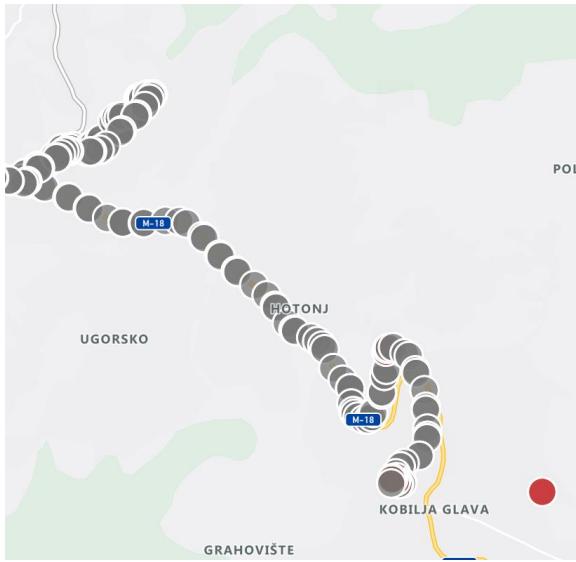


Fig. 6: Anomaly on a real map.

or not, so that the actual accuracy of the approach can be calculated.

Regarding the introduction of machine learning or artificial intelligence to this approach, two problems can appear, considering the data set used here. First one is that the data set is not balanced, because there are a lot more non-anomalies than anomalies, so that could be a problem for training and testing of machine learning. The second problem is that, in the data set, we do not have column that states whether the measurement is actually an anomaly or not, so the data set used for this paper is not appropriate to use in the case of machine learning, but some other data set, which has a column that says whether a measurement is an anomaly or not, would be appropriate.

REFERENCES

- [1] "A brief History of GPS In-Car Navigation", [Online, Accessed 23.05.2019] Available at: <https://ndrive.com/brief-history-gps-carnavigation/>
- [2] "Potential Problems with GPS Tracking", [Online, Accessed 23.05.2019] Available at: <https://www.autoalert.me.uk/problems-with-gps-tracking/>
- [3] Z. Liao, Y. Yu, and B. Chen, Anomaly detection in GPS data based on visual analytics, in VAST 10 - IEEE Conference on Visual Analytics Science and Technology 2010, Proceedings, 2010.
- [4] V. Patil, P. Singh, S. Parikh, and P. K. Atrey, GeoSClean: Secure Cleaning of GPS Trajectory Data Using Anomaly Detection, in Proceedings - IEEE 1st Conference on Multimedia Information Processing and Retrieval, MIPR 2018, 2018.
- [5] A. Zhang, S. Song, J. Wang, and P. S. Yu, Time series data cleaning: From anomaly detection to anomaly repairing, in Proceedings of the VLDB Endowment, 2017.
- [6] H. Wang, H. Wen, F. Yi, H. Zhu, and L. Sun, Road traffic anomaly detection via collaborative path inference from gps snippets, Sensors (Switzerland), 2017.
- [7] H. Moritz, Bulletin godsique, 1980.
- [8] R. Rodrguez, A. Mexicano, J. Bila, S. Cervantes, and R. Ponce, Feature extraction of electrocardiogram signals by applying adaptive threshold and principal component analysis, J. Appl. Res. Technol., 2015.
- [9] E. Zunic, S. Delalic, K. Hodzic, and Z. Tucakovic, "Innovative GPS Data Anomaly Detection Algorithm inspired by QRS Complex Detection Algorithms in ECG Signals," in 18th International Conference on Smart Technologies, IEEE EUROCON 2019, 2019.

- [10] E. Zunic, H. Hindija, A. Besirevic, K. Hodzic, and S. Delalic, Improving Performance of Vehicle Routing Algorithms using GPS Data, in 2018 14th Symposium on Neural Networks and Applications, NEUREL 2018, 2018.
- [11] Y. Quan, L. Lau, G. W. Roberts, X. Meng, and C. Zhang, Convolutional neural network based multipath detection method for static and kinematic GPS high precision positioning, Remote Sens., 2018.