Exploratory data analysis of stream data in sports medicine domain

Maja Vrancich
maja.vrancich@inf.uniri.hr
University of Rijeka, Department of Informatics
Rijeka, Croatia

ABSTRACT

In statistics, exploratory data analysis (EDA) is an approach to analyzing datasets to summarize their main characteristics, often with visual methods. A statistical model may optionally be used, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis-testing task. In this paper, we focus on analyzing stream data for the sports medicine domain, collected with a Polar M400 sports watch, with the aim of predicting the heart rate that the athlete must sustain on each race segment in order to achieve the best overall result (finishing time). We first provide an overview of data analyzing models for stream data. Next, we present related work, our data, and preliminary insights from visualizations after data preprocessing. Finally, we conclude the paper with a summary and future work.

KEYWORDS

Data analysis, stream data, time-series data, sports medicine, heart rate.

ACM Reference Format:
Maja Vrancich. 2019. Exploratory data analysis of stream data in sports medicine domain. In . ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The era of social fitness has drawn an increasing number of people to the world of distance running. In the United States alone, 18.1 million people registered for a race in 2018 (professional and recreational runners). [11] The participation of recreational runners in running races peaked in 2016 with a total of 9.1 million results, then declined to 7.9 million in 2018 [1].

Advances in GPS technology have enabled users to track outdoor activities like running, swimming, hiking, or cycling with their smartphones or GPS-enabled watches. Instead of merely gauging average pace, users can record their exact routes, as well as elevation changes, heart rate (HR) frequency, and other data. Additionally, online tracking and analysis sites such as Strava[1] enable users to document and share training history. Large sets of sport data are being accumulated in both training and race contexts. By studying the collected data, we can learn much about the nature of sport activity training and racing [12].

The explosive growth and widespread accessibility of digital health data have led to a surge of research activity in the fields of healthcare and data science. The conventional approach to health data management has achieved limited success, as it is incapable of handling the huge amount of complex data with high volume, high velocity, and high variety [6].

While studies have shown that machine learning methods can provide good performance in various healthcare applications for personalized disease diagnosis, medication, and treatments, it is still challenging to learn efficient patterns from heterogeneous healthcare data. Discovering hidden patterns in sequential data is still an open challenge, since it requires intelligent segmentation and clustering of the time-series data. For example, the time lapse between successive elements in patient records can vary from days to months, which may lead to suboptimal performance for traditional LSTM models. In addition, it is also difficult to interpret their performance, especially when the data is high-dimensional, which in turn limits the ability to design better architectures [20].

In sports such as cycling, the main variable for training planning and analysis is the power expressed by each pedal stroke, but in running, and particularly in trail running, there is no universally accepted method to measure the power output of the athlete. For that reason, the main variable we consider in this study is heart rate. Accurate heart rate monitoring is essential in fitness, training, and testing. Clapp and Little [4] showed that manual pulse palpation provides inaccurate results. The use of the electrocardiogram (ECG) or Holter monitoring is too costly and complex for athletes to use in the field. The first wireless heart rate monitor, the portable Polar PE 2000, was introduced in 1983. It consisted of a transmitter and a receiver. The transmitter could be attached to the chest using either disposable electrodes or an elastic electrode belt. The receiver was a watch-like monitor worn on the wrist. The wireless Polar heart rate monitoring method was developed in the Department of Electronics at the University of Oulu. In the beginning, the heart rate monitors were targeted for coaches and athletes to optimize the quality and efficiency of training. Soon, exercise scientists started to research the monitors and use them in their work. Today, the selection of heart rate monitors includes easy-to-use products for everyone interested in wellness, fitness, and health [13].

Sports watches had been one of the most important training tools for every amateur and professional athlete in the past, and after incorporating GPS technology, they became even more valuable. These watches have huge advantages over previous generations of devices, because they precisely measure characteristic training data. As a result, runners and cyclists do not need to use any special

[1] https://www.strava.com/

https://doi.org/10.1145/nnnnnnn.nnnnnnn

sensors for determining their speed, altitude, or duration of activity [9].

Selecting the most promising algorithm to model and predict a particular phenomenon is the main interest of temporal data forecasting. Forecasting (or prediction), similarly to other data mining tasks, uses empirical evidence to select the most suitable model for the problem at hand, since no modeling method may be considered the best [17]. Each tracked activity is saved into a file. Data mining could be applied to this collection of data, which would help athletes analyze their workouts, predict their further training activities, give advice about nutrition, etc. [8].

This paper is organized as follows. In Section 2 we go over related work. Next, in Section 3 we present our data and the results of EDA. Finally, we present future work and conclude the paper in Section 4.

2 RELATED WORK

Current state-of-the-art in data analysis for time-series data based on HR comprises diverse research, ranging from clinical decision-support systems, over heart disease predictions, to smart coaching. In this section, we briefly review existing work in the sports domain, which is closely related to our proposed problem.

Fister et al. [9] introduce a novel intelligent planning method for training sessions: training plans are computer-generated using the Bat Algorithm, according to reliable data obtained from sports watches.

Gronwald et al. [10], use detrended fluctuation analysis to assess heart rate correlation properties, examine the influence of exercise intensity on total variability and complexity in non-linear dynamics of heart rate variability.

Billat et al. [2] detect marathon asymmetry with a statistical signature. They tested marathon running performance, and revealed significant statistical features by analyzing speed time-series data recorded by 273 runners' GPS trackers. The combination of trend and asymmetry build up a statistical signature for the speed time-series, which is identical regardless of performance level, gender, or race profile.

Using data from the GPS-based Strava application, Marty [15] predicts the speed of individual riders, at specific times in the day, on a 2-mile segment of a popular cycing track using Ridge Regression model.

Jin [12] uses data from previous training runs, and applies four regression models to predict a run pace for a specific route or segment in future runs: basic linear regression, as well as Weighted, Ridge regression, and Lasso regression.

Clermont et al. [5] used wireless tri-axial accelerometer to classify competitive and recreational male and female runners (with greater than 80 % accuracy), as labelled by an objective group allocation method. Based on the features used in the support vector machine (SVM) models, they have shown that competitive runners have a more consistent and mechanically efficient running gait pattern than recreational runners.

Pharoah [18] develops and validates a new walking pace function using crowd-sourced GPS data. There are several functions hikers use to predict walking time based on elevation change or slope, the most popular of which is the Naismith function. Muazu Musa et al. [16] classified and predicted high (HCA) and low potential archers (LPA) from a set of physical fitness variables trained on a variation of k-NN algorithms and logistic regression.

Parmezan et al. [17] provides a systematic literature review of the last decade, identify state-of-the-art models for time-series prediction, and test those methods on 95 datasets. They conclude that SARIMA is the only statistical method able to outperform, but without a statistical difference, the following machine learning algorithms: ANN, SVM, and kNN-TSPI. However, such forecasting accuracy comes at the expense of a larger number of parameters.

3 DATA COLLECTION

The dataset used for this project was extracted from the Polar Flow API[2]. Polar Flow is a free online tool for planning and analyzing training, activity, and sleep, using data tracked with Polar devices.

The analyzed data was collected from a single athlete's Polar M400 sports watch, in the period between June 2017 and July 2019. In that time, 133 training sessions were logged in various sport profiles: 'running', 'trail running', 'hiking', 'orienteering', 'vertical sports wall climbing', 'watersports kayaking', 'strength training', 'mobility dynamic', 'mobility static', 'core', 'other outdoor', and 'other indoor'. The Polar M400 wrist watch is equipped with a GPS sensor, and paired with a chest strap heart rate monitor. Each training session's data is stored in a separate .json file. This data was preprocessed, and multiple features used for data visualization.

The data was parsed into two separate csv files: "summary data" (further reference MV01), containing a brief overview of each logged training session; and "seconds data" (further reference MV02), containing detailed information recorded for every second of a particular training session.

In outdoor activities, GPS related data is stored in intervals of one second: HR (heart rate), altitude, speed, distance and GPS position (longitude and latitude). This were the training sessions we are interested in because we want to find out how is one's heart responding to different terrain (slopes) and to progress over time due to training sessions done in the past.

In parsed files (MV01 and MV02), if some row of the file had missing data, it meant that it either wasn't done outside (no GPS related data like speed and altitude) or that heart rate strap haven't been worn (no HR data). There was no reason for keeping this rows and they were eliminated (dropped). In the end only training sessions of running, trail running and hiking were kept.

Table 1 lists the parameters recorded in MV01. The data in this file shows a general picture of the training sessions, and is used for extracting metadata about the person conducting the training sessions. In this case, VO_2 max, maximum and resting heart rate, aerobic and anaerobic threshold were changed only once. To increase data precision, the VO_2 max [19] test should be done periodically.

From summary data given in Figure 1, we can easily see the total distance and ascent done in training sessions of every month. When the distance bar is higher then the ascent bar, the training sessions were done on easier terrain (less altitude difference gained), and on steeper terrain in the opposite case. The longest distance was crossed in July 2018 (76.5 km), followed by May and June 2019 (55.3 and 54.5 km, respectively). Additionally, Figure 2 shows the average

[2] https://flow.polar.com/

heart rate by month. One might expect higher average HR in months of higher training intensity, but this is not the case. This data is not very useful or explanatory without information on how “hard” the training was (which is based on speed and altitude difference). That is why we moved to more informative data, aggregated in MV02.

Parameters used in MV02 are shown in Table 2. Data was collected from the same training session files, and again preprocessed.

We used only training sessions of running, trail running and hiking, and missing or incomplete data was dropped. Altitude difference was calculated over intervals of one second, and data aggregated in intervals of approximately 100 meters. Slope classification was added according to Barcelona Field Studies Centre [3] to each segment (see Table 3), along with ascent or descent labels, and the data was further grouped by slope and aggregated. We can see in Figure 3 that most training sessions were done on very gentle and gentle slope.

This data is going to be used in further modeling and predicting the best HR for each incline category of future races, based on provided GPS data on the particular track, and generated heart model of the athlete.

It is visible from Figure 4 that, while average HR has its own peaks (highest HR is achieved on races), average speed increases constantly over time, which means that athlete’s overall form is improving over time.

Table 1: Attributes of summary data file

Parameter	Description
startTime	Date and time at the beginning of the training session
stopTime	Date and time at the end of the training session
sport	Activity practised in given training session (for example: running, trail running, orienteering, stretching, hiking, core, etc.)
VO2max	VO_2 max is the maximal rate of oxygen uptake it is an important determinant of cardio-respiratory fitness and aerobic performance [19]
maximumHeartRate	Maximum heart rate – a variable that is editable in an athlete’s on-line profile (based on latest test)
restingHeartRate	Resting heart rate
aerobicThreshold	Aerobic threshold – the exercise intensity (HR) beyond which blood lactate concentration is no longer linearly related to exercise intensity, but increases with both exercise intensity and duration [14]
anaerobicThreshold	Anaerobic threshold – maximum steady-state lactate concentration [7]
distance	Overall distance crossed in given training session, measured in meters (every training session starts at 0 and counts meters continuously)
ascent	Overall ascent done in given training session, measured in meters
descent	Overall descent done in given training session, measured in meters
HRmin	Minimal HR measured during given training session
HRavg	Average HR measured during given training session
HRmax	Maximal HR measured during given training session

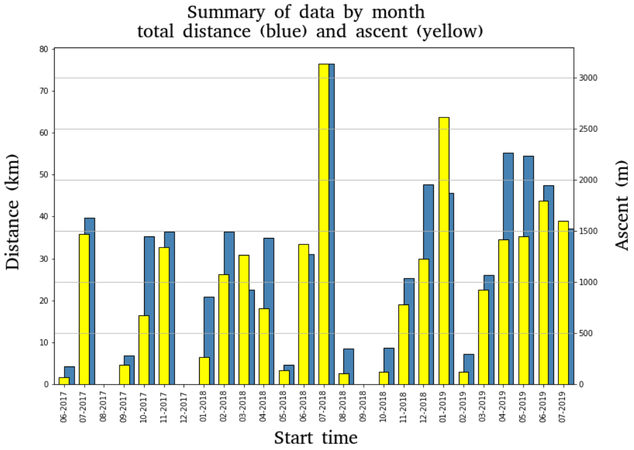

Figure 1: Summary of data by month; total distance (blue) and ascent (yellow)

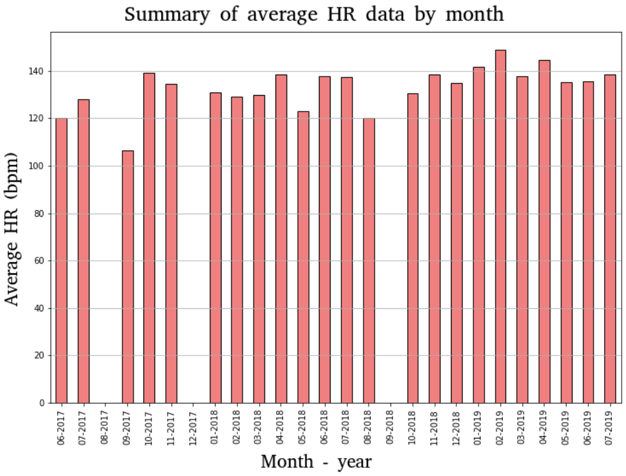

Figure 2: Summary of average HR data by month

4 CONCLUSION AND FUTURE WORK

In recent years, wearable smart watches and devices have become very popular and widely used, along with the data-collecting capabilities they provide. Different makers of sports watches are making

Table 2: Attributes of detailed data file

Parameter	Description
datetime	Exact time when sensors send data, every 1 second for every training session
f_num	File number, used for easier data editing
altitude	Altitude expressed in meters (exact altitude in given second)
heartrate	Heart rate frequency (e.g. 155)
speed	Moving speed in given second
distance	Distance, measured in meters (every training session starts at 0 and counts meters continuously; this parameter shows meters crossed from start of session to current point in time)
sport	Activity practised at given training session (for example: running, trail running, orienteering, stretching, hiking, core, etc.)

Table 3: Standard slope descriptors

Slope (%)	Approximate degrees	Terminology
0 - 0.5	0	Level
0.5 - 2	0.3 - 1.1	Nearly level
2 - 5	1.1 - 3	Very gentle slope
5 - 9	3 - 5	Gentle slope
9 - 15	5 - 8.5	Moderate slope
15 - 30	8.5 - 16.5	Strong slope
30 - 45	16.5 - 24	Very strong slope
45 - 70	24 - 35	Extreme slope
70 - 100	35 - 45	Steep slope
> 100	> 45	Very steep slope

efforts to improve their products, in order to make them more accurate, and thus more helpful to users wanting to lead healthy lifestyles and gain improvements in sports and personal fitness.

The ultimate goal of our research is building a model for predicting a runner's ideal heart rate for particular moments and segments of their training session or future race, developed by taking into account gradual fitness improvements through time (learned through monitoring regular training), and dependant on current position (slope and incline, distance, time elapsed). Combining this information with the GPS route provided for a particular training or race track, the HR model would provide a prediction of finishing time and best maximal HR on different sections of the route, to help the athlete adjust their pace throughout the running session.

The first step in future research is to apply the most appropriate data mining methods for time-series data analysis, e.g. SARIMA, Support Vector Machine and kNN-TSPI, in order to extract interesting and usefull knowledge from our data.

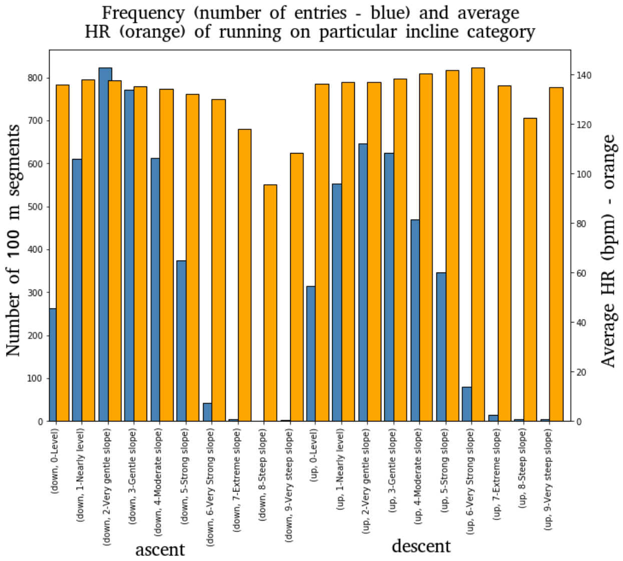

Figure 3: Frequency (number of entries - blue) and average HR (orange) of running on particular incline category

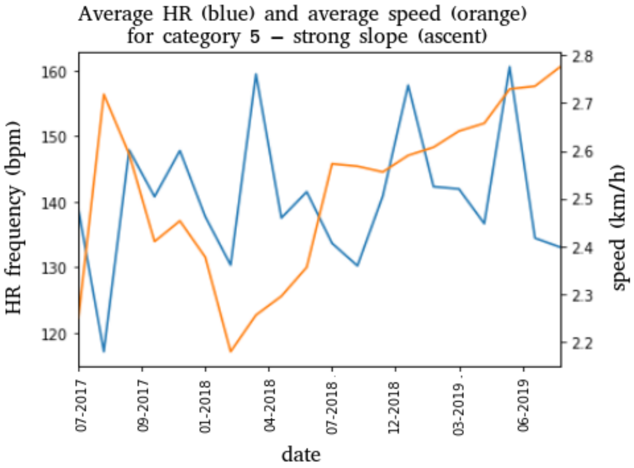

Figure 4: Average HR (blue) and average speed (orange) for category 5 – strong slope (ascent)

ACKNOWLEDGMENTS

This work has been fully supported by the University of Rijeka under the project number uniri-drustv-18-122.

REFERENCES

[1] Jens Jakob Andersen. 2019. The State of Running 2019. https://runrepeat.com/state-of-running

[2] Véronique Billat, Thomas Carbillet, Matthieu Correa, and Jean-Renaud Pycke. [n.d.]. Detecting the marathon asymmetry with a statistical signature. 515 ([n. d.]), 240–247. https://doi.org/10.1016/j.physa.2018.09.159

[3] Barcelona Field Study Center. 2019. Measuring Slope Steepness. https://geographyfieldwork.com/SlopeSteepnessIndex.htm

[4] JF Clapp and KD Little. 1994. The physiological response of instructors and participants to three aerobics regimens. *Medicine and science in sports and exercise* 26, 8 (August 1994), 1041—1046. http://europepmc.org/abstract/MED/7968422

[5] Christian A. Clermont, Lauren C. Benson, Sean T. Osis, Dylan Kobsar, and Reed Ferber. [n.d.]. Running patterns for male and female competitive and recreational runners based on accelerometer data. 37, 2 ([n. d.]), 204–211. https://doi.org/10.1080/02640414.2018.1488518

[6] Ruogu Fang, Samira Pouyanfar, Yimin Yang, Shu-Ching Chen, and S. S. Iyengar. [n.d.]. Computational Health Informatics in the Big Data Age: A Survey. 49, 1 ([n. d.]), 1–36. https://doi.org/10.1145/2932707

[7] Oliver Faude, Wilfried Kindermann, and Tim Meyer. [n.d.]. Lactate Threshold Concepts: How Valid are They? 39, 6 ([n. d.]), 469–490. https://doi.org/10.2165/00007256-200939060-00003

[8] Iztok Fister, Iztok Fister, Duan Fister, and Simon Fong. [n.d.]. Data Mining in Sporting Activities Created by Sports Trackers. In *2013 International Symposium on Computational and Business Intelligence* (2013-08). IEEE. https://doi.org/10.1109/iscbi.2013.25

[9] Iztok Fister, Samo Rauter, Xin-She Yang, Karin Ljubič, and Iztok Fister. [n.d.]. Planning the sports training sessions with the bat algorithm. 149 ([n. d.]), 993–1002. https://doi.org/10.1016/j.neucom.2014.07.034

[10] Thomas Gronwald, Olaf Hoos, Sebastian Ludyga, and Kuno Hottenrott. [n.d.]. Non-linear dynamics of heart rate variability during incremental cycling exercise. 27, 1 ([n. d.]), 88–98. https://doi.org/10.1080/15438627.2018.1502182

[11] Martin Fritz Huber. 2019. The Running Boom Isn't Going Anywhere. https://www.outsideonline.com/2393643/american-running-trends-survey-2019

[12] Tiffany Jin. [n.d.]. Predicting Pace Based on Previous Training Runs. ([n. d.]), 5.

[13] Raija M. T. Laukkanen and Paula K. Virtanen. [n.d.]. Heart rate monitors: State of the art. 16 ([n. d.]), 3–7. Issue sup1. https://doi.org/10.1080/026404198366920

[14] Theresa Mann, Robert Patrick Lamberts, and Michael Ian Lambert. [n.d.]. Methods of Prescribing Relative Exercise Intensity: Physiological and Practical Considerations. 43, 7 ([n. d.]), 613–625. https://doi.org/10.1007/s40279-013-0045-x

[15] Rachel Marty. [n.d.]. Predicting Bicycle Speeds on Mission Bay with Strava. ([n. d.]), 5.

[16] Rabiu Muazu Musa, Anwar P. P. Abdul Majeed, Zahari Taha, Siow Wee Chang, Ahmad Fakhri Ab. Nasir, and Mohamad Razali Abdullah. [n.d.]. A machine learning approach of predicting high potential archers by means of physical fitness indicators. 14, 1 ([n. d.]), e0209638. https://doi.org/10.1371/journal.pone.0209638

[17] Antonio Rafael Sabino Parmezan, Vinicius M.A. Souza, and Gustavo E.A.P.A. Batista. [n.d.]. Evaluation of statistical and machine learning models for time series prediction: Identifying the state-of-the-art and the best conditions for the use of each model. 484 ([n. d.]), 302–337. https://doi.org/10.1016/j.ins.2019.01.076

[18] Paul Pharoah. [n.d.]. Development and validation of a new walking pace function using crowd-sourced GPS data. ([n. d.]). https://doi.org/10.1101/188417

[19] Niels Uth, Henrik Srensen, Kristian Overgaard, and Preben K. Pedersen. [n.d.]. Estimation of VO2max from the ratio between HRmax and HRrest the Heart Rate Ratio Method. 91, 1 ([n. d.]), 111–115. https://doi.org/10.1007/s00421-003-0988-y

[20] Qinghan Xue, Xiaoran Wang, Samuel Meehan, Jilong Kuang, Alex Gao, and Mooi Choo Chuah. [n.d.]. Recurrent Neural Networks based Obesity Status Prediction Using Activity Data. ([n. d.]). arXiv:1809.07828 http://arxiv.org/abs/1809.07828