Author(s) anonymized for blind review stage

Lap sector segmentation using discrete fourier transformation and geospatial alignment for inter- and intra-athlete workout file comparison

Author address info removed for blind review stage

Abstract

In this paper we introduce a workflow for automatic lap detection in Global Positioning System (GPS) workout files followed by subdivision in distance-based sectors. First, Discrete Fourier Transform (DFT) power spectrum analysis is used to detect the number of completed laps in the athletes' workout files. Subsequently, when the number of completed laps is available, a geospatial neighborhood based search procedure detects the lap split points in the workout file. Combination of the number of laps and the consecutive lap boundaries enables sector based geospatial alignment. This is achieved by a mechanism, aligning the base course with laps of several workouts, based on a combination of cumulative lap distance and geospatial distance between sector start- and end-point. Geospatial alignment allows straightforward sector based performance comparison of the recorded workouts of circular mass-start sports events. As mentioned, workout laps are mapped on a base course, which is usually a hand-drawn GPS trace, or a GPS file offered by the race organizers. The base course is further subdivided in a number of fixed-length sectors, allowing a more detailed comparison between a set of laps. Course sector start- and endpoints are matched with the closest points in the participants' laps extracted from their uploaded workout files. As these matching points are timestamped, the elapsed time between both points serves as a completion time of the sector in question. The proposed workflow enables inter- and intra-athlete lap comparison and is providing additional insights such as: an overview of the sectors where the biggest differences were made; the *n* hardest sectors to complete based on lap time or the engagement score of a sector based on the average and the standard deviation of duration.

Keywords

GPS analysis, Fourier Transform, sector time comparison, geospatial alignment

Introduction

The current advancements in development and deployment of Global Positioning System (GPS) technology has made GPS receivers widely available and affordable (Kumar et al., 2002).

GPS aided technology (such as smartphones and wearables) are widely used by sports enthusiasts to track and evaluate their progress. Modern GPS head units and watches are nowadays capable of recording time stamped geographical data (latitude, longitude, elevation, distance) combined with additional sensor data (heart rate, power, cadence and temperature) and produce a workout file containing the recorded geographical and sensor data. Interpretation and evaluation of these files is usually performed by athletes or coaches. Visualization and interpretation are facilitated by one of the many software packages that calculate performance metrics and map or graph the data from GPS workout files. These tools are usually centred around coaching and tracking the fitness progress of an athlete with aids such as (semi-)automated workout planning and automated analysis of executed workouts.

Most of the aforementioned software packages are mainly focused around evaluation of sensor data. GPS data is commonly used with the sole purpose of visually liking the sensor data with the location where it was recorded. Only a few tools are focussed around direct geospatial analysis of workout files. Strava, an online app that tries to build a social network of like minded sports enthusiasts, is one of the software packages that offers thorough geospatial analysis of uploaded GPS logs from fitness devices. Quite a while ago, they introduced the concept of timed GPS segments, which allows users to compete on predefined road segments without the need to physically take part in a planned race or group activity.

To the best of our knowledge, the utilised algorithms to search and match segments in GPS are entirely proprietary. This makes athletes dependent on the features and comparisons offered by the companies of the segment matchers. In this paper, we want to introduce an open and high-performing way to geographically analyze circular GPS data gathered on mass start events such as cyclocross races or the final laps of a road world championship. In the remainder of this paper, we propose a workflow to automatically detect the number of laps, followed by a segmentation of the file into lap segments, concluded by the alignment of the laps, allowing inter- and intra-athlete lap time comparison. We verify the introduced techniques on GPS files originating from uploaded files of competitive cyclocross races.

Methods

As already discussed in the previous section, we'll try to make a contribution in the domain of segment matching in large geospatially annotated datasets. The process of lap extraction and the subsequent matching can be subdivided into three different intermediate steps.

The first step consists of the extraction and preprocessing of the geospatially data, which is usually stored in structured file formats such as Flexible and Interoperable data Transfer (fit), Training Center XML (tcx) or GPs eXchange format (gpx). To facilitate both inter- and intra-athlete comparison we resampled the timestamped geospatial records to a sampling interval of 1 Hz as it simplifies the subsequent processing steps of the lap matching process. As processing libraries do exist in any of the popular programming languages, the first step of our matching process is a rather straightforward job. The upsampling is achieved by linear interpolation of both the geospatial coordinates (Equation 1) and the recorded sensor data. Downsampling isn't required as the maximum sampling frequency of the commercially available GPS head units and watches is capped at 1 Hz. Although it might not be ideal for racing or exercising, the utilization of standalone GPS receivers, sampling at 10 Hz, might potentially deliver more accurate results, as they were proven to offer better insights in instantaneous speed differences (Gløersen et al., 2018).

$$t \ \epsilon \ [0, T], \qquad p_t = (lat_t, lon_t)$$

$$\boldsymbol{then}: p_i \ \in [p_1, p_2] \ \Leftrightarrow \ p_i + \frac{t_i - t_1}{t_2 - \ t_1}(p_2 - p_1)$$

Equation 1 : *the interpolation of the latitude (lat) and longitude (lon) for a point p_i between timestamped geospatial points p_1 and p_2*

In the next step of the process the number of laps was derived from the time stamped (1 Hz) coordinates of the circular GPS workout files. As opposed to the first step this procedure involved less obvious techniques. As we are trying to detect circular laps, having the same start and end point, a large number of geospatial coordinates will recur in a predictable way. Figure 1 shows the recurrence of longitude coordinates recorded during a Belgian cyclocross race. This is a very interesting finding as this allows us to reduce three-dimensional GPS data (latitude, longitude and timestamp) to a two-dimensional (either latitude or longitude and timestamps) data structure. Within this regard, GPS data is very similar to an audio signal which is physically described as a disruption of local air pressure at frequencies within the audible range. The challenge within audio processing is often to decompose an audio signal. Audio is often a combination of multiple harmonic sound waves. In audio processing Discrete Fourier Analysis (DFT) is a technique to detect the harmonic content of such synthesized sound waves. In brief, DFT is a technique which allows to find the “repetivity” within a temporal signal. As our timestamped GPS coordinates are temporal and also recurrable over time, Fourier analysis can indicate the number of laps a rider completed of a certain course. In Figure 2, the power spectrum is shown for the longitude coordinates of Figure 1. With the help of Fourier analysis the longitude signal was decomposed into a set of sinusoids of different frequencies. Consequently, the power spectral density graph (as shown in Figure 2) showed how the power of the total signal, which is the variation in longitude in our example, is distributed across the detected sinusoids. The frequency with the highest power is the most likely candidate for the number of laps. Figure 2 illustrates that according to the power spectrum Fourier analysis that the frequency of 10 (laps) is the most powerful across the spectrum, which is confirmed by a longitude pattern that is repeating ten times over elapsed time in Figure 1.

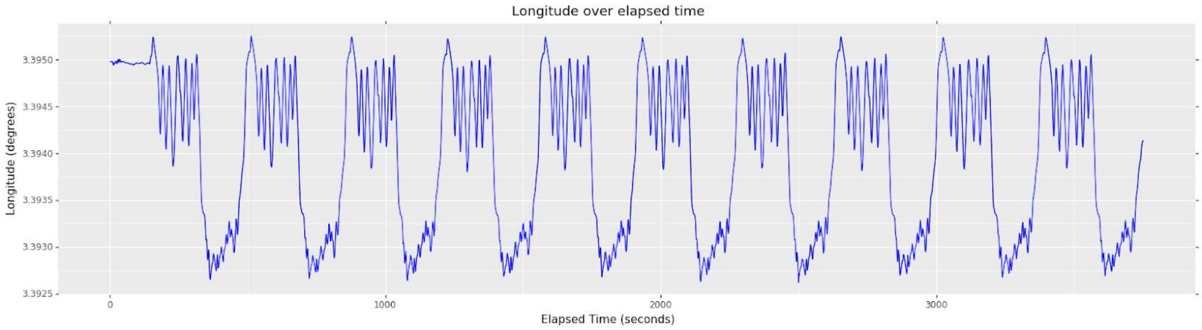

Figure 1 : *Longitude plotted over elapsed time, with ten repetitions of the same longitude pattern.*

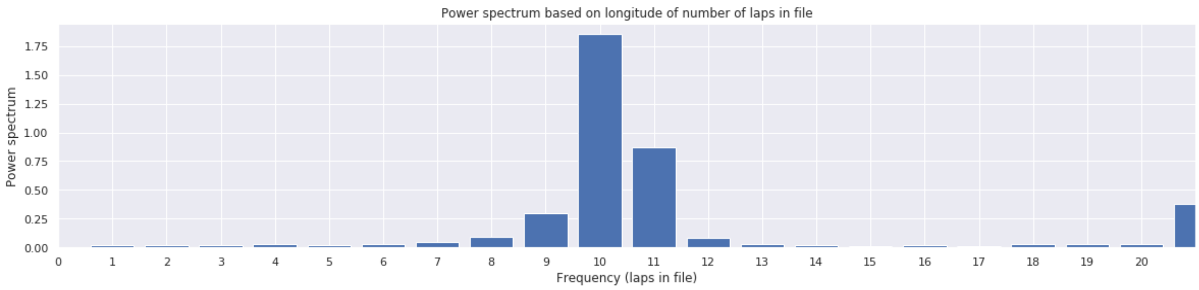

Figure 2 : *Power spectrum Fourier analysis for the longitudes of Figure 1 : Longitude plotted over elapsed time, with ten repetitions of the same longitude pattern.*

Now that we have a number of geospatially repeating patterns from our GPS coordinates over time we can find the exact lap boundaries. A lap boundary is defined as the point on the course where the previous lap is completed and a new lap is started. For the exact starting point the point at which the repetitive pattern is starting has been chosen, which in cycling races is often the start- and finish-area.

With these aspects in regard, the lap boundaries can be calculated by investigating all the coordinates of the athlete’s workout file and look for a point on the course that is recurring at a similar frequency as the number of laps. This is achieved by looking at the start of the GPS file, provided a small offset to take into account the fact that most athletes start their recording devices slightly before the start of a race, and find a coordinates which is reoccuring the next lap. To find this subsequent passage through the start point the algorithm searches for the closest point around the ideal lap point (pt_{calc}, Eq. 2). This point is defined by starting point plus total workout time (T, Eq. 2) divided by the number of laps (n, Eq. 2) detected by the DFT analysis. The algorithm looks within an interval around the ideal lap point. This interval was defined at fifteen percent of the total workout time plus or minus the timestamp of the ideal lap point. This optimization greatly reduces the total number of comparisons. Equation 2 summarizes the algorithm to extract the lap points from the workout file when it's provided by the number of completed laps.

$$T = T_{total,workout} - offset_{start+end}$$

$$pt_{calc,i} = pt_{act,i-1} + \frac{1}{n} \cdot T$$

$$pt_{act,i} = \min\left[dist(pt_{act,i}, pt_j)\right] \quad \forall\, j \in \left[pt_{calc,i} - \frac{15}{100} \cdot T;\ pt_{calc,i} + \frac{15}{100} \cdot T\right]$$

Equation 2 : *the calculation of the next lap split point pt_{act} , if provided with the previous lap point $pt_{act,i-1}$, workout time T and a DTW-based calculation of number of laps n.*

To make sure that the first calculated next point is close enough to the course start point an accuracy radius was set. If the next point isn't within the radius of the candidate lap point we move the index of the starting point ahead and repeat the process described in Eq. 2 with the new start point pt_{act} set to $pt_{start,\, i+1}$.

$$\boldsymbol{if}\ dist\left(pt_{start,i}, pt_{next}\right) > radius \rightarrow pt_{start,i} = pt_{start,i+1}$$

Equation 3 : *Start point quality control process: ensures that the starting point for the lapping procedure is close enough to the detected next point.*

Now the number of laps and their exact lap boundaries were extracted from the workout files we switch to the lap alignment procedure that matches the coordinates of these laps (inter- and intra-workout). The alignment process is achieved by further subdivision of the laps in sectors, which are fixed parts of the course. All the sectors of the course are having the same distance (i.e. the *sector length*). Alignment of the detected laps with the sectors of a hand drawn base course is achieved based on start- and end-point of the sectors. External circumstances (e.g. deep forest, bad weather, …) which are resulting in an often rather inaccurate recorded workout are imposing the need to introduce a twofold procedure to get the best possible sector point alignment between an often inaccurate workout file and a more precise hand drawn GPS course. The first step iterates over all the sector split points and matches course sector split points to lap split points entirely based on elapsed distance in both paths (see Pseudocode 1).

```
lap_index ← 0
Sector_lap_splits ← []
for sector_index, elapsed_distance in course_sectors
      while elapsed_lap_distance[lap_index] < elapsed_distance
          lap_index ← lap_index + 1
    End while
      sector_lap_splits.push(lap_index)
End for
```

Pseudocode 1: *First iteration to match sector split points of course and lap coordinates based purely based on elapsed distance.*

This first iteration which is $max(O(n_{sectors}), O(n_{lap\_points}))$ produces a set of points which are possible lap sector splits in the workout files. As mentioned, the variable accuracy of GPS data, necessitates an additional iteration which aims to further fine-tune the splits in the lap data. This iteration mainly uses a weighted sum of both elapsed distance difference and haversine distance between course and lap sector split points (*Eq. 4*). The search space to find the best matching point is limited by the points of the lap which are within an interval limited by the previous and next sector lap split point (calculated in the first iteration, *pseudocode 1*). This extra iteration is $O(n_{lap\_points})$ which makes this two step lap matching algorithm $O(n_{lap\_points})$.

$$d_{sector_j, pt_i} = \min \sqrt{\left(sector_{length} \cdot abs\left[1 - \frac{elapsed_{course,j}}{elapsed_{lap,i}}\right]\right)^2 + distance\left(pt_{lap,i}, pt_{sector,j}\right)^2}$$

$$\forall\, i\ \in \left[pt_{sector,j-1} + \frac{1}{4} \cdot n\,, pt_{sector,j+1} - \frac{1}{4} \cdot n\right]$$

$$n \leftarrow number\ of\ points\ between\ [pt_{sector,j-1}, pt_{sector,j+1}]$$

Equation 4 : *Second iteration which takes closest lap point i to the course point at sector at index j ($pt_{sector,\,j}$ is the split point of sector j in the lap points) based on a distance ($d_{sector_j,\,pt_i}$) weighting elapsed distance and haversine distance between course and laps sector point*

The last and rather straightforward step is to get the elapsed lap sector times based on the points in the lap aligned with the sector split points. The time required to complete a sector of the course is calculated by the entire lap time needed to reach the point at the end of that sector minus the time to reach the starting point of that sector, which is also the end of the previous sector.

Results

The proposed methodology to extract the number of laps, extract separate lap data from the entire workout file and the matching of fixed-distance lap sectors with those of the base course are ultimately offering sector times for all the available activity files of participating athletes on a circular course. The travelled distance-based sector times are allowing a direct comparison between sector times, both intra- and inter-athlete. The former, as researched by Hopkins et Al. (2001), can show interesting insights in the consistency of an athlete over the duration of a race. In sports or sport disciplines which involve a technical component, sector times are considerably influenced by obstacles or technically demanding zones of the course. The calculated sector times of an athlete across his or her completed laps of the course can be further examined based on median, mean or standard deviation of the completion times of a sector. This type of examination is illustrated in Figure 3. The race was a cyclocross round of a pro general classification series in the U23 category in Belgium. For those not familiar with cyclocross, this is basically a race over repeating laps consisting of a mixture of well-rideable and technical, muddy or steep sector. The course is tackled with slightly modified road bicycles (wider and knobby tires and a more robust and higher frame geometry). The rider in the example completed a total of 6 laps on a roughly 9 minute during lap. The fifth lap (in purple) was plotted against the area between the first- and third-quartile (Q1 & Q3) sector values for the rider's lap times. The more narrow the surface between Q1 and Q3, the more consistent our rider was at a given sector. The rider finished well in the top ten of the race, which is fortifying Hopkins' (2001) statement that the more consistent the athletes' sector times the higher their chances to ride a good race resulting in a good end ranking. The rider's lap interquartile-range distances of sector times for the rider in Fig. 3 are well between a couple of seconds. Furthermore, the graph is also showing that at sector 15 and 29 the rider is riding remarkably slower as usual. We can only guess for the exact reason for this relatively slower sector times, but a non-exclusive list of possible factors are: the fatigue building up during the is race impacting the performance, weather conditions changed during the race or maybe due to a technical mistake of the rider. Extra info for our example is that both sectors were highly technical and became more muddy during the race as ruts formed and a rain shower started in the second half of the race.

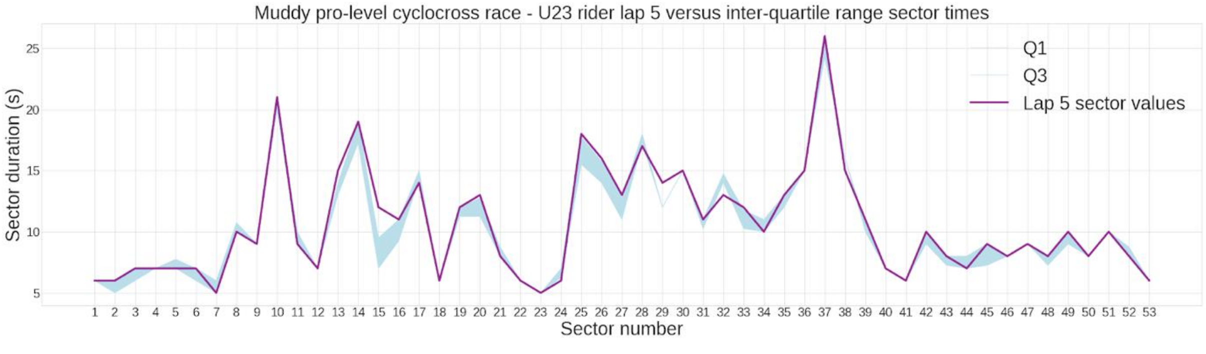

Figure 3 : *Visual analysis of sector times for lap 5 (purple line) of an U23 rider participating in an elite cyclocross race compared with Q1-Q3 range of sector times of the athlete.*

The previous case study investigated the intra-athlete analytical possibilities of our suggested mechanism. Another interesting investigation possibility is the inter-athlete analysis for a certain event. With the help of some first order statistics such as median, and quartiles we can already get a basic understanding of the faster and slower sector of our course. Figure 4 shows a plot of these statistical measurements for the same muddy course as in the intra-athlete case study. For this example we used a total of 9 available race workout files originating from three different categories (U23, Women Elite, Men Elite). Each category has a different race duration (50 minutes, 45 minutes and 60 minutes respectively) providing us with 53 time measurements for each of the fixed-length sectors of the course. In the graph of Fig. 4 we can clearly see 3 distinctive peaks. After further examination, two out of the three correspond with sectors in which the riders had to dismount their bikes and run up a steep and muddy hill. The other peak was an off-camber muddy sector on which a small mistake could also result into a forced dismount of the rider. Furthermore, we could also observe larger variability (light blue area between Q1 and Q3, Fig. 4) at the more technical sectors of the course. A pro cyclocross course often starts and ends at a paved sector of road. This pattern could also be recognized in the graph: the start and end sectors have consistently lower sector times as those in the middle.

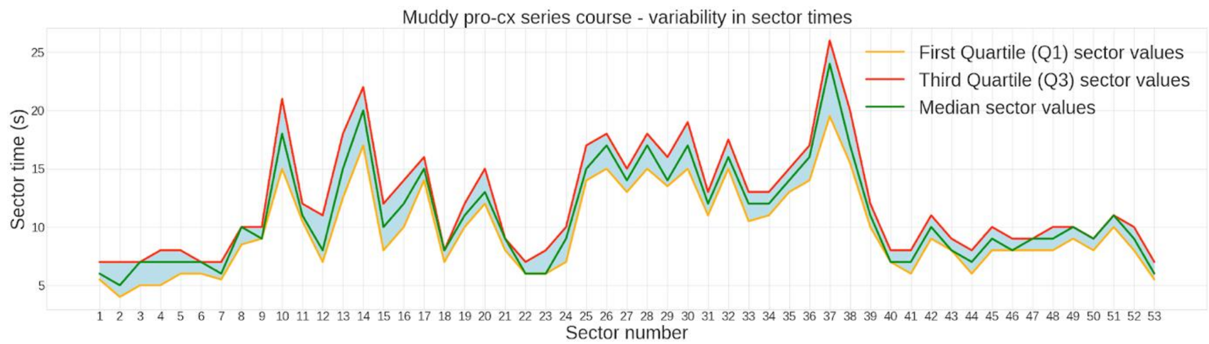

Figure 4 *: Variability in sector times of the same muddy course as in Fig. 2 & 3 of various athletes (9) of different categories (3). The more difficult the sector (higher sector) time, the more variability.*

As illustrated, sector times and their first order statistics are offering interesting insights into the track layout and might perhaps even reveal some of the track conditions. This arises the possibility to propose a sector engagement scoring that could potentially find it benefits in various use cases. Broadcasters of the race could for instance use this information to position extra equipment at the most engaging sectors or the race organizers could use engagement scorings of previous editions or of other races to build a highly engaging course. In Equation

5, we propose such a scoring mechanism that is solely relying on the statistics based on the available sector times.

$$score_i = \mathbf{ln}\left(\frac{med_i}{max - min}\right) \cdot med_i + \left(1 - \frac{med_i}{max - min}\right) \cdot \overline{med}$$

$$\textbf{with}: \quad med_i \leftarrow median\ of\ the\ athletes'\ sector\ times\ for\ sector\ i$$

$$max \leftarrow maximum\ of\ all\ the\ median\ values\ of\ the\ sector\ times$$

$$min \leftarrow maximum\ of\ all\ the\ median\ values\ of\ the\ sector\ times$$

$$\overline{med} \leftarrow average\ of\ all\ the\ median\ values\ of\ the\ sector\ times$$

Equation 5 : *Calculation of a sector engagement scoring based weighting of a sector's median value, overall maximum and minimum sector time, median of the median sector values.*

The engagement scoring for our muddy course of the previous examples could be observed in Figure 5. It is obvious that the running and off-camber sectors which we discovered in Figure 4 are also ranked as the most engaging by the mechanism. Additionally, the aforementioned start finish area is also ranked as less engaging in comparison with the off-road sectors.

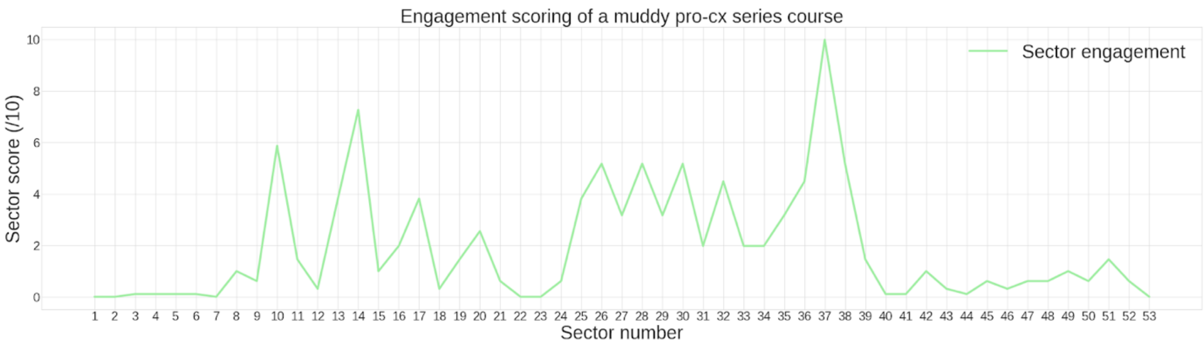

Figure 5 : *Suggested engagement ranking (based on Eq. 5 and rescaled from 1 to 10) of the various sectors of the muddy course of the previous examples (Fig. 2-4).*

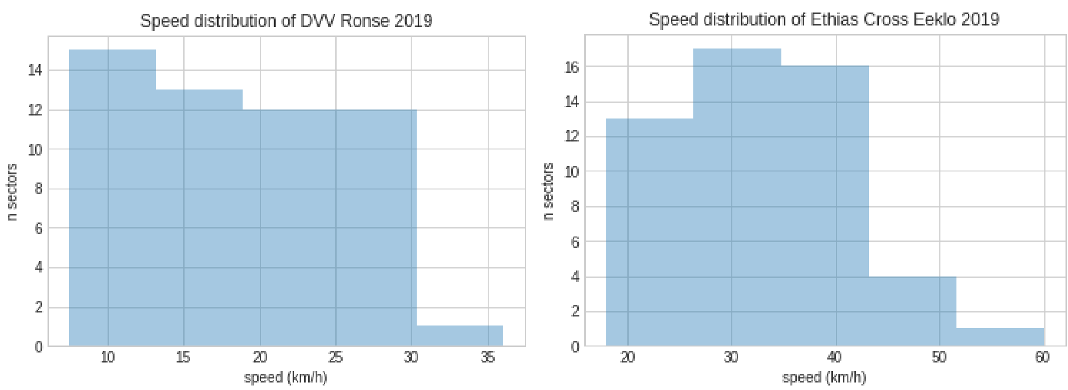

Figure 6 *: Outlier sector times with GPS traces of the sector in question which is illustrating that GPS accuracy is very low around that area.*

Finally and as suggested by Hurst et al. (2016), GPS data can also be used to characterize the course and course conditions. As Figure 6 is demonstrating, median sector times can give additional insights in the type of course. In downhill racing this principle was used to compare different downhill tracks in the same bike park. For cyclocross this method can also be used, but in cyclocross it is perhaps more interesting to compare multiple tracks with different weather conditions. Figure 6 is demonstrating this principle, the left graph is showing the sector speed distribution of a muddy, hilly race. The right graph is an example of a fast and dry course from the beginning of the Belgian cyclocross season. The first graph is clearly and consistently showing slower sector speeds.

Discussion

The mechanism and its applications are a good starting point for further development of a lap comparison toolkit for circular mass-start events. Combination of the sector times with additional metadata about the sectors of the course could provide more insights in the course and course circumstances. A big step towards this goal would be the registration of some of the track characteristics such as track width, surface type, surface condition and possible obstacles on a fixed-length sector level. Combination of sector times and the additional information about the sectors in question might even allow direct comparison between sectors of different courses or editions of a race. This would enable similarity based matching of sectors across different courses when a dataset consisting of sector times and its accompanying metadata is provided.

Likewise, and provided that sufficient workouts from the same category (i.e. U23, Women Elite, Men Elite, …) are available we could provide a performance ranking mechanism, similar to the engagement scoring mechanism which is giving a rider a score for each sector based on his ability on that specific sector compared to the other riders in his/her category.

A final future consideration is focussed around the fact we are dealing with real-life GPS data and that it is sometimes challenging to provide accurate insights when the provided data isn't accurate. Acosta and Toloza (2012) mention that the GPS accuracy is between 10 and 15 meters for 95% of the time. As a matter of fact, this is already a considerable margin, but combined with the typical characteristics of cyclocross courses this can sometimes offer impactful accuracy issues. Cyclocross in particular is a sport in which riders have to ride through variable terrain, with courses that are often laid out in a very condensed area, containing a lot of tight turns. All of the above is resulting in GPS reception and accuracy which is not always as good as it could or should be for accurate position tracking. These circumstances often result in a relatively low amount of logged coordinates around that area and ultimately cause the mechanism to provide sector times which are rather inaccurate. From the other side, in the future this can also serve as an inaccurate sector sample detection mechanism, as the calculated sector times often fall outside the human possibilities or are well outside the range of the other durations for that sector.

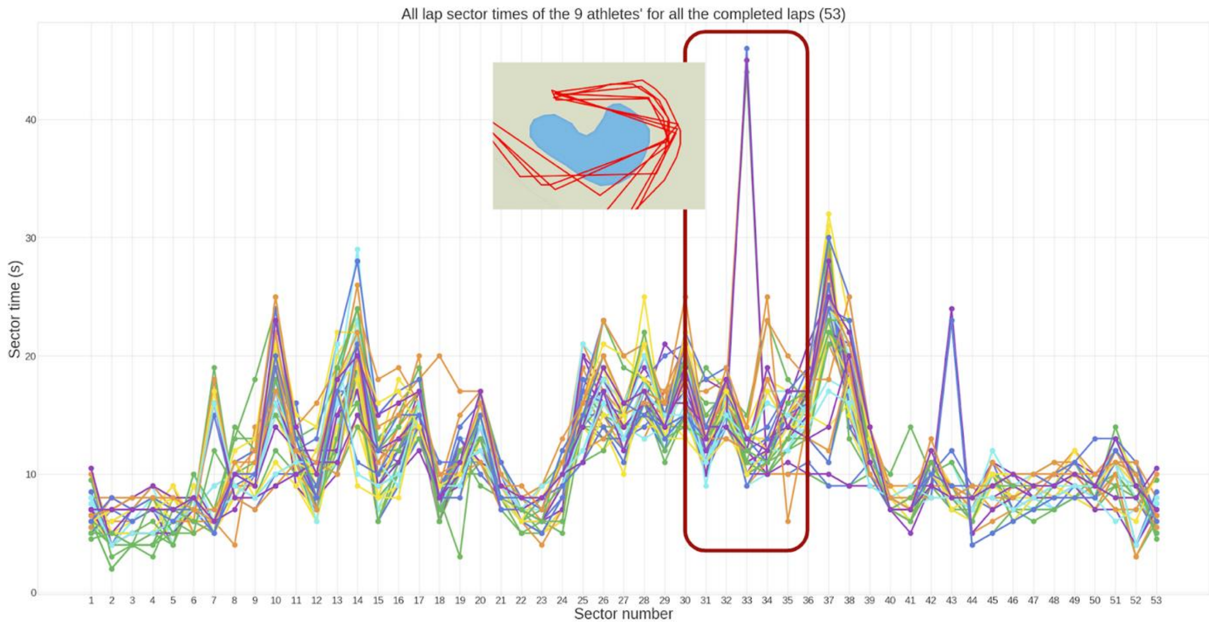

Figure 7 : *Outlier sector times from GPS traces. The overlaid map containing the sector traces of the outliers is illustrating that GPS accuracy was very low around that area.*

Practical applications

The presented methodology in this paper is directly applicable in endurance sports in which athletes are using a GPS-enabled recording device. The test results discussed in this paper were conducted on a collection of Belgian cyclocross races of different categories. Cyclocross is a cycling discipline that is very suited for our approach as riders repeat the same lap between six and ten times (depending on the exact course distance).

The ability to isolate challenging and engaging sectors is also promising for other disciplines and even other types of endurance sports. Currently, we are extending the proposed GPS-based sector extraction methodology to also support non-circular road racing. We already performed some initial tests in which we successfully extracted and analyzed riders' sector completion times of some of the famous Belgian cobbled climbs. Analysis of climbs or cobbled sectors during road races has the added challenge of variable sector lengths. This extra challenge requires an alternative engagement ranking algorithm to allow comparison of variable length sectors based on the competitors' sector times. Figure 8 shows the ten most interesting sectors of the 2019 Tour of Flanders. In hindsight, the final race outcome was decided by the combination of the well-known duo of the “Oude Kwaremont” (third passage) and the “Paterberg” (second passage), which is also observable in our engagement rankings in Fig. 8 as they are respectively the fourth most and the most engaging sectors.

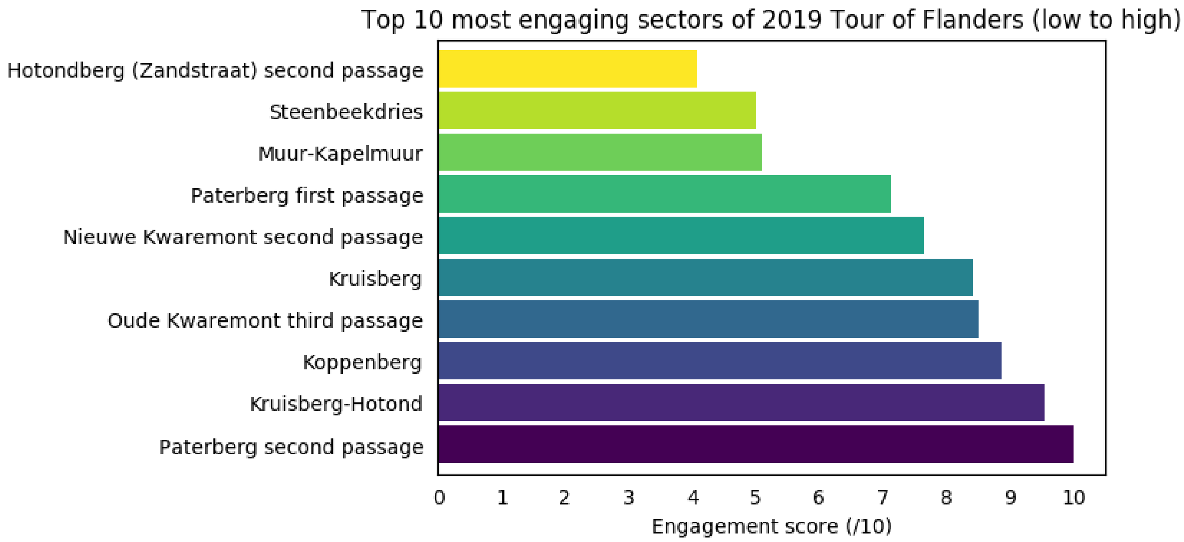

Figure 8 *: Sector engagement score based on riders' files of the 2019 Tour of Flanders*

Lastly, collection and linkage of other metadata about the riders (e.g. age, nationality, weight or height) or the sectors themselves (e.g. wind and weather circumstances or average gradient) could further extend both the scope of application and the abilities of our ranking mechanism. For road races, factoring in the wind direction and speed into the absolute sector times could provide additional insights into the engagement of a sector on a specific day (with given weather circumstances). Likewise, for technical, off-road disciplines such as cross-country or downhill mountain biking, extra info about the track (e.g. type of surface, condition of the surface or amount of incident sunlight) combined with the sector timings of the practice runs could potentially reveal the sectors on which the difference will be made on race day.

Conclusion

In this paper we presented a mechanism to extract the number of laps and split the workout in different laps. The extracted laps were further subdivided in fixed length sectors which we used, based on the lap data of different athletes and the timestamps of the matching sector split points, to calculate sector completion times. The availability of sector times per athlete per lap allows both inter and intra-athlete comparison and was found to be beneficial for engagement scoring, checking a rider's consistency during the race and even provide a basic idea of the course type and/or conditions.

Acknowledgments

This research is funded by the imec ICON project DAIQUIRI (Agentschap Innoveren en Ondernemen project nr. HBC.2019.0053).

References

Acosta, N., & Toloza, J. (2012). Techniques to improve the GPS precision. *International Journal of Advanced Computer Science and Applications, 3*(8). doi:10.14569/ijacsa.2012.030820

Berg, R. E., & Stork, D. G. (2005). *The physics of sound*. Pearson Prentice-Hall.

Hopkins, W. G., & Hewson, D. J. (2001). Variability of competitive performance of distance runners. *Medicine & Science in Sports & Exercise, 33*(9), 1588-1592. doi:10.1097/00005768-200109000-00023

Gløersen, Ø., Kocbach, J., & Gilgien, M. (2018). Tracking Performance in Endurance Racing Sports: Evaluation of the Accuracy Offered by Three Commercial GNSS Receivers Aimed at the Sports Market. *Frontiers in physiology*, *9*, 1425. https://doi.org/10.3389/fphys.2018.01425

Hopkins, W. G., & Hewson, D. J. (2001). Variability of competitive performance of distance runners. *Medicine & Science in Sports & Exercise, 33*(9), 1588-1592. doi:10.1097/00005768-200109000-00023

Hurst, H. T., & Atkins, S. (2006). Power output of field-based downhill mountain biking. *Journal of Sports Sciences, 24*(10), 1047-1053. doi:10.1080/02640410500431997

Kumar, S., & Moore, K. B. (2002). *Journal of Science Education and Technology*, *11*(1), 59-80. doi:10.1023/a:1013999415003

Equations and pseudocode

Illustrations