IDENTIFYING PACING PROFILES IN 2000 METRE WORLD CHAMPIONSHIP ROWING

A PREPRINT

Dani Chu
Simon Fraser University
dani_chu@sfu.ca

Renny Doig
Simon Fraser University
rennyd@sfu.ca

Ming-Chang Tsai
Canadian Sport Institute Pacific
mtsai@csipacific.ca

Jack Davis
Simon Fraser University
jackd_@sfu.ca

Ryan Sheehan
Simon Fraser University
rsheehan@sfu.ca

January 19, 2020

ABSTRACT

The pacing strategy adopted by athletes is one of the major determinants of successful performance during timed competition. Various pacing profiles are reported in the literature and its potential to predict a winning performance depends on the mode of sport. However, in 2000 metre rowing, the definition of these pacing profiles has been limited by the use of minimal data. Our aim is to objectively identify pacing profiles used in World Championship 2000 metre rowing races in a reproducible and transparent way. To do this the average speed and stroke rate (SR) for each 50 metre split for each boat in every race of the Rowing World Championships from 2010-2017. This data was scraped from www.worldrowing.com. This dataset has been made publicly available (https://github.com/danichusfu/rowing_pacing_profiles) in order to further the field of rowing research in general. Pacing profiles are then determined by using k-shape clustering on the average boat speeds at each 50 metre split. Finally, clusters are described using boat and race descriptors to draw conclusions about who, when, and why each pacing profile was observed. A multinomial logistic regression is then fit to test whether variables such as boat size, gender, round, or rank are associated with pacing profiles. Four pacing strategies (Even, Positive, Reverse J-Shaped, and U-Shaped) are identified from the clustering process. Boat size, round (Heat vs Finals), rank, gender, and weight class are all found to affect pacing profiles.

***K*eywords** rowing, pacing profiles, time series clustering

1 Introduction

Across "closed-loop" design sports, competitions where athlete(s) attempt to complete a set distance in the shortest time [Abbiss and Laursen, 2008], there have been different pacing strategies that have been identified. Most of these pacing strategies have been defined in running and cycling races and an attempt has been made to define these strategies in 2000m rowing ([Garland, 2005]; [Kennedy and Bell, 2003]; [Muehlbauer and Melges, 2011]; [Muehlbauer and Melges, 2011]). However, these attempts approach the problems in a different manner and come to different conclusions. We attempt to standardize the definition of pacing profiles in rowing by using more granular data than other attempts.

Determining optimal pacing profiles can be done using ergometric data [Kennedy and Bell, 2003] or by using observational data from actual competitions ([Garland, 2005]; [Muehlbauer and Melges, 2011]; [Muehlbauer et al., 2010]).

[Kennedy and Bell, 2003] used simulated rowing and training results to suggest that there were different optimal race profiles for different genders. They found that a constant pacing profile was optimal for men and an all-out profile was optimal for women.

[Garland, 2005] used observational data from the 2000 Olympics, 2001 World Championship, and 2001 & 2002 British indoor Rowing Championship competitions. The analysis found that when using four time splits measured every 500 metres that men and women show no difference in their observed pacing strategies. Garland eliminated races that showed signs of slowdowns from the analysis. Then [Muehlbauer et al., 2010] as well as [Muehlbauer and Melges, 2011] used the same type of split time data to model pacing profiles. In 2010 they found that gender, round of race (whether race was in qualifying heat or the final race for the category), size of boat, coxed, and scull did not affect pacing strategies for the 2008 Olympics. In 2011 they had a different finding that indicated that round of race affected pacing profiles in World Championship races between 2001 and 2009. They performed these analyses by fitting linear quadratic models to the four time splits.

The classification of pacing profiles has often been approached by fitting linear models to split times. We believe that using more granular data describing a boat's speed throughout the race will be able to paint a better picture of how the boat is performing throughout the race. We also believe that using a clustering technique to classify similarly shaped speed curves together will provide a novel approach to defining pacing profiles. There is a large body of literature in clustering and the area of longitudinal clustering is growing. In sports specifically, model-based clustering has been used to cluster player trajectories in basketball [Miller and Bornn, 2017], football [Chu et al., 2019], and soccer [Gregory, 2019]. [McNicholas et al., 2012] used a model-based clustering approach that uses mixtures of multivariate t-distributions with a linear model for the mean and a modified Cholesky-decomposed covariance structure to cluster gene expressions. Additionally, [Kumar and Futschik, 2007] used a soft clustering technique to cluster the shapes of microarray data. Finally, using UCR time-series datasets [Chen et al., 2015], a collection of datasets that has been collected to test clustering techniques and improve the clustering techniques that are published, [Paparrizos and Gravano, 2016] developed the k-shape clustering technique for time series data.

We match the clustering results to pacing profiles described in racing literature. We then determine which race factors are related to the use of a pacing profile and which are not.

To do so we present the following:

1. A data set (https://github.com/danichusfu/rowing_pacing_profiles) with data from the Global Positioning System (GPS), Media Start List, and Race Results from World Championships from 2010 to 2017. Additionally, we have included the code needed to scrape this data for future years and replicate our process of scraping and extracting future data.
2. A novel approach of classifying pacing profiles for boats in 2000m rowing.
3. A model to understand the effects of race factors on the use of pacing profiles.

2 Methods

2.1 Athletes and event

Data was gathered from www.worldrowing.com for the average speed and stroke rate (SR) at each 50 metre split for each boat in every race of the Rowing World Championships from 2010-2017. This includes both lightweight and heavyweight races, men, women, and mixed races, and all other race descriptors. Additionally, data was extracted that described the boats. Their finishing place, lane, country, discipline, category, and more were collected. In Table 1 we present the number of boats by year.

2.2 Data Analysis

Data was initially filtered to eliminate races with GPS errors where the reported average speed is lower than the true average speed, with an unreported average speed at any of the split measurements (at every 50 metres), average speed less than 2 metres per second, with boats that received "Did not Starts", "Did not Finishes" or "Exclusions". This reduced the number of boats' races from 9264 to 8170. To determine pacing profiles raw speeds at each split are often compared to the mean speed of a boat throughout the race [Garland, 2005]. So we define $x_{i,j}$, as the speed at split i for boat j and normalize to get $y_{i,j}$, where

$$y_{i,j} = \frac{x_{i,j} - \bar{x}_j}{\sigma_j} \tag{1}$$

Table 1: Number of Boats from each World Championship

Year	Men	Women	Total
2010	498	250	748
2011	911	467	1378
2012	765	424	1189
2013	759	387	1146
2014	776	466	1242
2015	1112	632	1744
2016	235	112	347
2017	772	383	1155

This is useful because the magnitude of the speed has been normalized and we can now compare the pacing profile of an eight to that of a single while accounting for the fact that their speeds will have different magnitudes.

With the boats' races cleaned and speeds standardized to the same scale we can begin the clustering process. The idea is to group speed curves of similar shape together.

In k-Shape clustering a new distance method, called "Shape-based distance (SBD)", and a new method for computing centroids are used. When SBD is evaluated against other distance metrics such as Dynamic Time Warping, it reaches similar error rates on the UCR datasets but with shorter computation times. The k-Shape algorithm is implemented in the `dtwclust` package [Sarda-Espinosa, 2018]. In its implementation it normalizes the columns to the same scale. So it takes $y_{i,j}$ defined in 1 and transforms it to $z_{i,j}$ defined as

$$z_{i,j} = \frac{y_{i,j} - \bar{y}_j}{\sigma_i}.$$

A k-Shape algorithm therefore functions very similarly to the k-means algorithm [Lloyd, 1982] in that the method uses iteratively defined clusters to minimize within-cluster distance.

We fit a multinomial logistic regression with pacing profile as a dependent variable on the boat size, race placement in a heat or final, discipline, gender, and weight class variables. To fit the logistic regression model we removed all boats from mixed-gender races and adaptive races due to lack of data. This left 8054 boats on which to train the model on. Below we report the odds ratio for each variable in the model. An effect is determined to be a statistically significant if the p-value from the Wald z-test is smaller than 0.05 divided by 39 (the number of significance tests we are doing). This is a test with $\alpha = 0.05$ and a Bonferroni correction applied so that our global Type I error rate is 0.05.

3 Results

3.1 Type of pacing profiles

In other races of fixed distance like cycling and running six pacing profiles have been defined [Abbiss and Laursen, 2008]. The six profiles are "negative", "all-out", "positive", "even", "parabolic-shaped", and "variable pacing".

A negative-split pacing profile is defined by an increase in speed across splits (which result in smaller relative split times as the race progresses) and is often used in middle-distance events.

An all-out profile is used when the it is believed that energy reserves are best distributed at the start of the race. This is commonly found in shorter events like the 100 metre sprint.

A positive pacing profile is one where the athletes' speed decreases through each split in the event. This is found often in swimming, where the diving start allows athletes to reach their maximum speed quickly.

Even pacing profiles are categorized by a relatively small portion of the race spent in the acceleration phase and the majority of the race at a constant pace.

There are three pacing sub-strategies for Parabolic-Shaped pacing profiles. J-Shaped, Reverse J-shaped, and U-shaped. In general these strategies follow a parabolic shape where the middle of the race sees the lowest relative speeds. In the U-shaped strategy, the start and end of the race see the same relative speed. The J-Shaped strategy has a greater relative speed at the end of the race while the Reverse J-Shaped profiles have a greater relative speed at the start of the race.

The last profile mentioned is "Variable Pacing". It is a strategy that is used to adapt to changing conditions in the race course, like uphills and downhills in cycling.

Table 2: Summary of clusters for $k = 4$

Cluster	Size	Average Distance Within
1	1951	0.1233
2	2277	0.0956
3	2548	0.1104
4	1444	0.1620

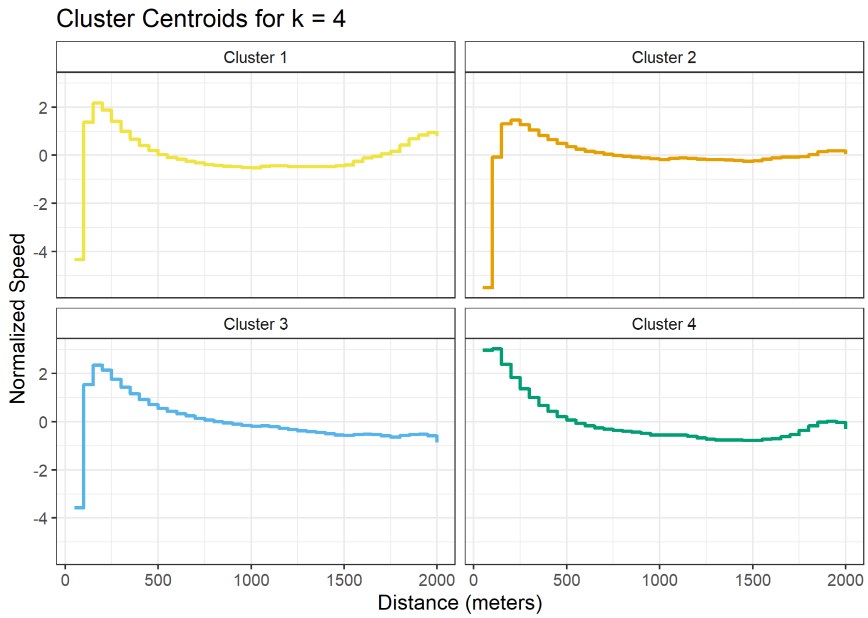

Figure 1: Cluster Centroids for k-Shape Clustering with 4 Clusters

3.2 Comparison of pacing profiles

We performed k-Shape clustering for $k = 3, 4$, and 5. We found that $k = 4$ gave us the most distinct shapes. The k-Shape clustering algorithm converges, which means there is an iteration of the algorithm where cluster memberships do not change, for our given seed. The size of the clusters and average distance between rows in each cluster are reported in Table 2.

Next to understand what shapes of clusters were found we plot the centroids for each cluster in Figure 1. We can see that while the centroids are similar, as expected in an all-race average, there are distinct features that separate them. The centroids are plotted with respect to the normalized speed by race $(y_{i,j})$. Again, this is so that we can identify the shape of the pacing curve without the effect of magnitude that size of boat, weight class, and other variables would affect.

We will now name the clusters based on the definitions given by [Abbiss and Laursen, 2008].

Cluster 1 is defined by a slow acceleration to a moderate peak velocity, a slow middle section, and a final sprint that almost reaches peak velocity. This agrees with the definition of the U-Shaped Pacing profile.

Cluster 2 is defined by a slower acceleration, a smaller peak velocity, and a low variance in speed throughout the rest of the race. This agrees with the definition of the Even Pacing profile.

Cluster 3 is defined by an acceleration to top speed in the first 150 metres and a decline in speed for every proceeding split. This agrees with the definition of the Positive Pacing profile.

Cluster 4 is defined by a quick acceleration to a higher peak velocity, a slower middle portion of the event, and finally a faster push to the finish. This agrees with the definition of the Reverse J-Shaped Pacing profile.

Table 3: Odds ratio changed by each variable holding all others constant. Statistically significant entries are bolded.

	Positive	Reverse J-Shaped	U-Shaped
Intercept	0.8972	0.7654	1.4110
Size: One-person (baseline)	–	–	–
Size: Two-person	**0.4795**	**0.3802**	**0.6796**
Size: Four-person	**0.1272**	**0.1360**	**0.1607**
Size: Eight-person	**0.0356**	**0.0757**	**0.0354**
Round of Race: Final (baseline)	–	–	–
Round of Race: Heat	**1.8120**	1.037	**0.5397**
Race Placement: 1st Place (baseline)	–	–	–
Race Placement: 2nd Place	0.8631	1.020	1.2070
Race Placement: 3rd Place	1.0780	1.3260	**1.4980**
Race Placement: 4th Place	1.3580	**1.6040**	**1.6000**
Race Placement: 5th Place	**1.7600**	**1.9280**	**1.2420**
Race Placement: 6th Place	**3.1620**	**3.1600**	**1.2050**
Discipline: Sculling (baseline)	–	–	–
Discipline: Sweep	**1.8140**	1.2010	**1.9660**
Gender: Men (baseline)	–	–	–
Gender: Women	**1.8830**	**1.6830**	**1.6630**
Weight Class: Lightweight (baseline)	–	–	–
Weight Class: Open	**1.4320**	**1.5230**	**1.2840**

3.3 Pacing profiles and race factors

To explain how to interpret these numbers we will use the boat size variable as an example. The first thing to note is that there is no results reported for the "Even" pacing profile. This is because it is used as our baseline level. Additionally, there is no estimates for single sculls. This is because the single sculls are the baseline for the boat size variable.

The odds that eights would follow a "Positive" pacing profile is 0.03 times as large as the odds that single sculls would follow an "Even" pacing profile holding all other variables constant. We can see that since all odds ratios for the eights are less than 1 that an eights is most likely to exhibit an "Even" pacing profile.

Holding all other variables constant, the "Positive" pacing profile is nearly 2 times more likely to be used in a heat than a final.

A given boat's placement in the race seems to have an effect on the selection of pacing profile. The baseline in this case is boats that came in 1st place. The question is whether this would affect how the boats would pace themselves. There is no significant difference between pacing profiles chosen by 1st place and 2nd place boats. 3rd place boats have a similar distribution but are more likely to have a "U-Shaped" pacing profile. 4th, 5th and 6th place boats have a wildly different distribution of odds of following a given pacing profile.

A boat's discipline is either called a "Sculling" or a "Sweep". "Sculling" describes a boat where rowers use 2 oars and "Sweep" describes boats where rowers have only one oar each. We see that "Positive" and "U-Shaped" pacing profiles are more likely in "Sweep" boats than "Sculling" boats.

Women were statistically less likely to follow "Even" pacing profiles when accounting for all other variables included in the model.

The "Open" weight class also saw a different distribution of pacing profiles compared to the "Lightweight" class after accounting for the other variables. Holding the other variables constant the "Positive" and "Reverse J-Shaped" pacing profiles were more likely to be used.

4 Discussion

4.1 Type of pacing profiles

The bigger the boat the more likely one was to observe an even pacing profile. This is most likely because it takes a lot of inertia for the bigger boats to get moving. Once a bigger boat reaches it's race speed there is not much variance in speed. It's harder for an eight to go fast and slow.

It was noted above that the "Positive" pacing profile is nearly 2 times more likely to be used in a heat than a final. This would make sense as boats that are in heats are more likely to want to conserve their energy for their future races. Boats will often race the first half of the race as if it's the final and then reassess if they should back off to conserve energy for future rounds. They do so based on where they rank and how far they are from the leader. This behaviour propagates from the slowest boats to the fastest. Once the fastest boats are ahead (usually by some margin) and they see the slower boats slow down, they will simply cruise as well; hence contributing to high odds of using positive racing strategy Interestingly, in rowing there is a rule that disqualifies boats that are not "trying their hardest" throughout the race [FIS, 2019].

The placement in a race was also found to be statistically significant for the last 3 places. One explanation could be that in most races the first 3 boats are the ones to qualify for the next race. As discussed above, once you have secured your placing (especially in heats) you begin to conserve energy. Third and fourth place boats are usually battling for a qualifying spot. So the top 2 boats will often be similar to each other and the bottom 2 boats will often be similar to each other. Unfortunately looking for an interaction between these the race placement and the round of the race would require more data than we have available so we leave this for further investigation.

5 Conclusion

Our approach makes an important contribution to the current literature. We provide an objective, data-driven approach to quantifying racing strategy. Previous analyses have been done through a subjective quantification. This approach uses a complex time series clustering method to characterize racing strategies. With these clusters, we developed a model which allows inference to be made about these racing strategies in relation to other factors present during a race. The granularity of the data we provide is what allows the methods we have presented to make accurate classifications. Furthermore the granular data collected has been made available to the public so that future analyses may be performed with similar accuracy.

Our biggest contribution is that we quantify racing strategies objectively based on data. Previous definitions have been done descriptively or with very small amounts of data. The added granularity of our data set allows for more accurate clarification of pacing profiles.

6 Acknowledgements

We would like to thank Dr. Dave Clarke for organizing the partnership with CSIP, Chuck Rai for his incredible help with scraping data, and Lucas Wu and Kevin Floyd for their consultations on statistics and rowing strategies. We would also like to thank Ron Yurko, Sam Ventura, Rebecca Nugent, and the Carnegie Mellon Sports Analytics Club for hosting the reproducible research competition that pushed us to make our work reproducible and available to the public. We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) where they support Dani Chu with the Alexander Graham Bell Master's Scholarship.

References

[FIS, 2019] (2019). Fisa rule book.

[Abbiss and Laursen, 2008] Abbiss, C. R. and Laursen, P. B. (2008). Describing and understanding pacing strategies during athletic competition. *Sports medicine (Auckland, N.Z.)*, 38:239–52.

[Chen et al., 2015] Chen, Y., Keogh, E., Hu, B., Begum, N., Bagnall, A., Mueen, A., and Batista, G. (2015). *The UCR Time Series Classification Archive*.

[Chu et al., 2019] Chu, D., Reyers, M., Thomson, J., and Wu, L. Y. (2019). Route identification in the national football league. *Journal of Quantitative Analysis in Sports*, 0(0).

[Garland, 2005] Garland, S. W. (2005). An analysis of the pacing strategy adopted by elite competitors in 2000 m rowing. *British Journal of Sports Medicine*, 39(1):39–42.

[Gregory, 2019] Gregory, S. (2019). Ready player run: Off ball run identification and classification. In *Proceedings of the 2019 Barca Sports Analytics Summit*.

[Kennedy and Bell, 2003] Kennedy, M. D. and Bell, G. J. (2003). Development of race profiles for the performance of a simulated 2000-m rowing race. *Can J Appl Physiol*, 28(4):536–546.

[Kumar and Futschik, 2007] Kumar, L. and Futschik, M. (2007). Kumar l, futschik e.. mfuzz: a software package for soft clustering of microarray data. bioinformation 2: 5-7. *Bioinformation*, 2:5–7.

[Lloyd, 1982] Lloyd, S. P. (1982). Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28:129–137.

[McNicholas et al., 2012] McNicholas, P., Sanjeena, and Subedi (2012). Clustering gene expression time course data using mixtures of multivariate t-distributions. *Journal of Statistical Planning and Inference*, 142:1114–1127.

[Miller and Bornn, 2017] Miller, A. C. and Bornn, L. (2017). Possession sketches : Mapping nba strategies. In *Proceedings of the 2017 MIT Sloan Sports Analytics Conference*.

[Muehlbauer and Melges, 2011] Muehlbauer, T. and Melges, T. (2011). Pacing patterns in competitive rowing adopted in different race categories. *The Journal of Strength & Conditioning Research*, 25.

[Muehlbauer et al., 2010] Muehlbauer, T., Schindler, C., and Widmer, A. (2010). Pacing pattern and performance during the 2008 olympic rowing regatta. *European Journal of Sport Science*, 10(5):291–296.

[Paparrizos and Gravano, 2016] Paparrizos, J. and Gravano, L. (2016). k-shape: Efficient and accurate clustering of time series. *SIGMOD Rec.*, 45(1):69–76.

[Sarda-Espinosa, 2018] Sarda-Espinosa, A. (2018). *dtwclust: Time Series Clustering Along with Optimizations for the Dynamic Time Warping Distance*. R package version 5.5.0.