Defining Pacing Profiles in World Championship 2000m Rowing

An Exploration through k-Shape Clustering

Dani Chu[1], Ryan Sheehan[1], Dr. Ming-Chang Tsai[2,3] & Dr. Jack Davis[1]

1. Simon Fraser University 2. Canadian Sport Institute Pacific 3. University of Toronto

Contact Information:

Ryan Sheehan
ryan_sheehan@sfu.ca

Dani Chu
dani_chu@sfu.ca
danichusfu.github.io

Dr. Ming-Chang Tsai
mtsai@csipacific.ca

Dr. Jack Davis
jack_davis@sfu.ca

Abstract

The pacing strategy adopted by an athlete(s) is one of the major determinants of successful performance during timed competition. Various pacing profiles are reported in the literature and its potential to precede a winning performance depends on the mode of sport. However, in 2000 meter rowing, the definition of these pacing profiles has been subjective and there is a need to be more objective with the definition. Our aim is to objectively identify pacing profiles used in World Championship 2000 meter rowing races in a reproducible transparent way. To do this the average speed and stroke rate (SR) for each 50 meter split for each boat in every race of the Rowing World Championships from 2010-2017 was Scraped from www.worldrowing.com. This data is made publicly available 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 meter split. Finally, clusters are described using boat and race descriptors to draw conclusions about who, when and why each pacing profile was observed. To do this a Multinomial Logistic Regression is fit with Bonferroni corrections used on the tests of significance 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. This novel approach of using clustering is able to objectively define four strategies used in 2000 meter rowing competitions.

Keywords: rowing, pacing profiles, k-Shape clustering, multinomial logistic regression, race analysis.

Introduction

- In "closed-loop" sports, pacing profiles, also known as speed curves, are well defined. [1]
- Our objective is to objectively identify pacing profiles used in World Championship 2000 meter rowing races.
- The literature of pacing profiles in 2000 meter rowing does not agree on which factors affect a boat's pacing profiles or what those pacing profiles are.
- With objectively defined pacing profiles we can identify in which situation boats exhibit different pacing profiles.
- To do so we present the following contributions to 2000m rowing analysis:
  1. A github repository with Global Positioning Sytem (GPS), Media Start List, and Race Results data from World Championships from 2010 to 2017 in an easy to use comma seperated value (csv) file. Scripts needed to scrape this data for future years and replicate our process of scraping and extracting the current data are provided.
  2. A novel approach of classifying pacing profiles for boats in 2000m rowing.
  3. Which race factors affect and do not affect the use of a given pacing profile during an event based on the fit of a Multinomial Logistic Regression.

Data Collection

- The availability of easily usable data in rowing is limited. Some studies have used split time results to analyze pacing profiles [2] [3] [4] , but there was been limited use of the publicly available, all be it harder to extract, GPS race data.
- Our data collection process involved writing a bash script to scrape all GPS PDFs, along with Media Start List PDFs and Race Results PDFs from every round of every category of the specified World Championships. This process takes roughly 25 minutes to run. In total there were 5321 working PDF files from the 8 years of World Championships.
- We removed any race that did not have complete information available online. There were 9264 boats' races with 131 variables available to analyze. Some of the variables include the average speeds and strokes per minute for each 50 meter split, race descriptors and boat descriptors.

Data Filtering and Pre Processing

We remove any boat that has a unreported average speed at any of the 40 split measurements (taken every 50 meters) and any boat that saw reported average split speeds less than 2 meters per second. Additionally, we removed any 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 [2]. So we define $x_{i,j}$, as the speed at split i for boat j and normalize to get $y_{i,j}$.

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

This is useful because the magnitude of the speed has been normalized and we can now compare the pacing profile of an 8 person boat to that of a 1 person boat while accounting for the fact that their speeds will have different magnitudes.

k-Shape Clustering

We use the k-Shape method developed by Paparrizos and Gravano for clustering time series data [5] and implemented in the **dtwclust** R package [6].

We performed k-Shape clustering for k = 3, 4, and 5. We found that k = 4 gave us the most distinct shapes. . To understand what shapes of clusters were found we plot the centroids for each cluster wutg respected to the normalized speed by race ($y_{i,j}$) in Figure 1. We can identify the shape of the pacing curve without magnitude. Which should size of boat, weight class and other variables would affect.

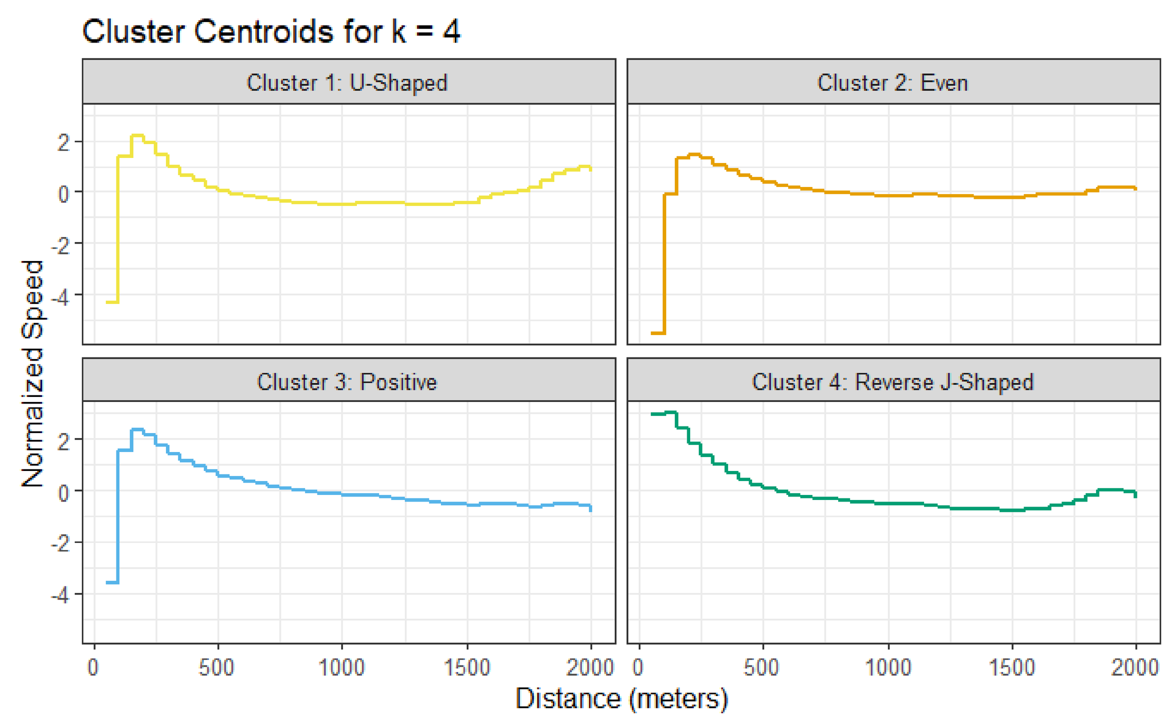

Figure 1: Cluster Centroids for k-Shape Clustering with 4 Clusters. See reasoning for names below.

Identifying Pacing Profiles

We will now name the clusters based on the definitions given by Abbiss for pacing profiles used in other closed loop sports [1].

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 coordinates 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 aligns with the definition of the Even Pacing profile.

Cluster 3 is defined by an acceleration to top speed in the first 150 meters and a decline in speed for every proceeding split. This fits 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 matches the definition of the Reverse J-Shaped Pacing profile.

Modelling Usage of Pacing Profiles

A multinomial logistic regression was fit with pacing profile as a dependent variable on the boat size, race placement, round (heat or final), discipline, gender and weight class variables. To fit the logistic regression model we removed all boats from mixed gender races, junior races, and adaptive races due to lack of races. This left 7306 boats to train the model on. Below we report the odds ratio for each variable in the model in Table 1. An effect is determined to be statistically significant if the p-value from the Wald z-test is smaller than $\frac{0.05}{27} = 0.000185$ (where 27 is the number of significance tests we are doing). This is a test with $\alpha = 0.05$ and a Bonferonni Correction applied so that our global Type II error rate is 0.05.

	Positive	Reverse J-Shaped	U-Shaped
Intercept	0.81	0.70	1.28
Size: 2 person	**0.48**	**0.41**	**0.74**
Size: 4 person	**0.12**	**0.16**	**0.17**
Size: 8 person	**0.03**	**0.09**	**0.03**
Heat or Final: Heat	**1.99**	1.13	**0.56**
Race Placement: 2nd Place	0.88	1.03	1.19
Race Placement: 3rd Place	1.11	1.32	**1.60**
Race Placement: 4th Place	**1.45**	**1.60**	**1.65**
Race Placement: 5th Place	**1.96**	**2.04**	1.31
Race Placement: 6th Place	**3.46**	**3.26**	1.20
Discipline: Sweep	**1.81**	1.07	**1.91**
Gender: Women	**1.85**	**1.62**	**1.71**
Weight Class: Open	**1.33**	**1.65**	1.12

Table 1: Change in odds ratio for each variable and pacing profile relative to the Even pacing profile. Bold values are statistically significant at a 95% global significance level (using a Bonferonni Correction).

Size of Boats

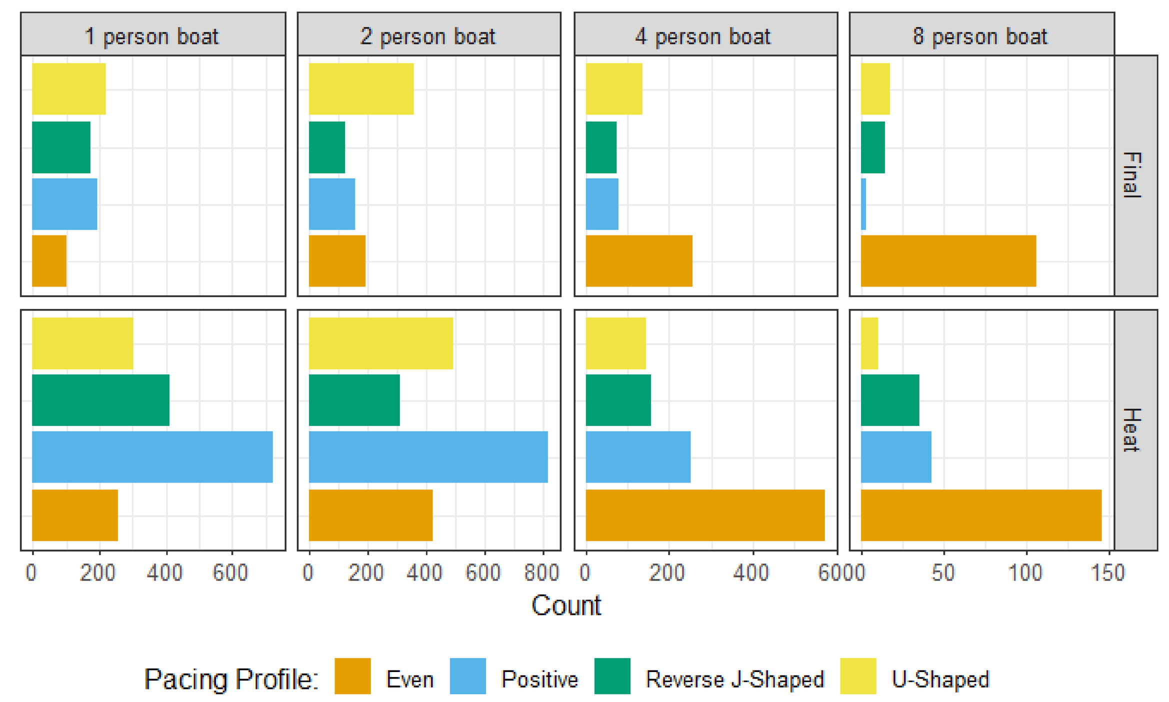

Figure 2: Pacing Profiles split by Size of Boat.

We clustered on a normalized boat speed for the race so it is interesting that we pick out size as a factor that affects pacing profile. This effect can be seen in Figure 2. One could argue that it is a different sport when there is a different number of athletes in the boat. An 8 person boat requires rhythm and coordination across all 8 members. On the other hand a boat with only 1 athlete has no one else to coordinate with. This gives them more control to change pace at will.

Heats vs Finals

The format for the World Championships has a Final (Final A) for placing the top 6 boats of the tournament and other Finals races (Final B, Final C...) to place the remaining boats.

A Positive pacing profile is used more often in Heats compared to Finals across all boat sizes. Specifically for 1 person boats we can also see that the proportion of U-Shaped pacing profiles we observe in Finals is larger compared to Heats.

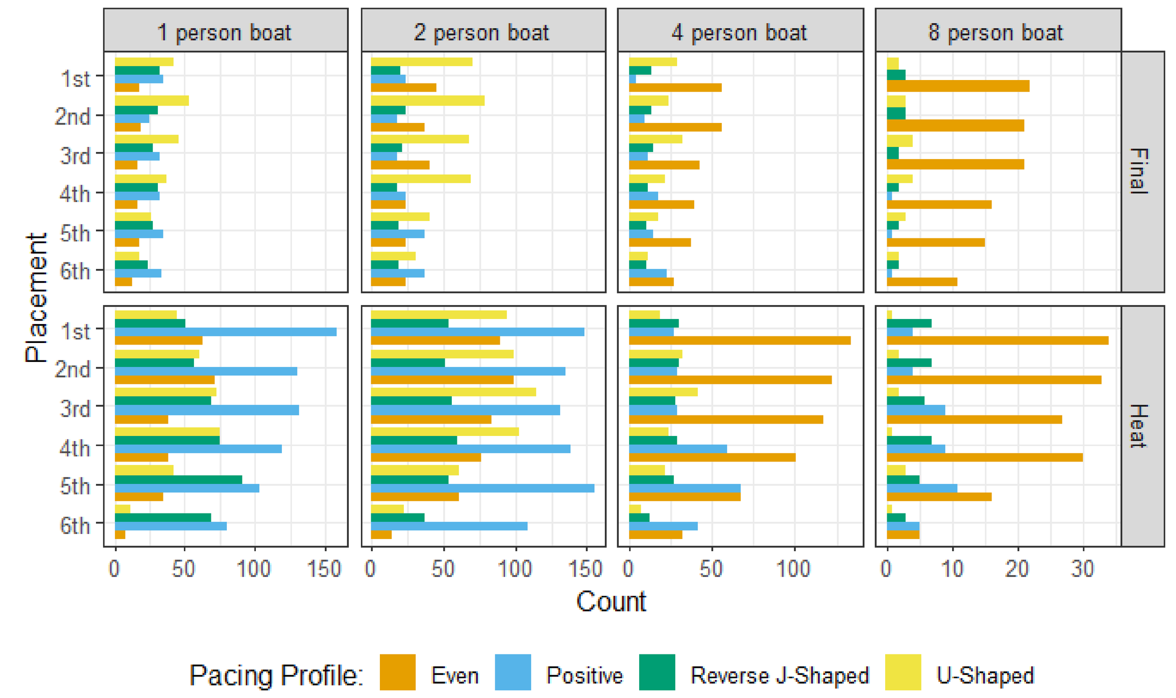

Figure 3: Pacing Profiles split by Size of Boat.

This leads us to propose a couple of new hypotheses about these 1 person boats.

1. The pacing profile of a boat is affected by their performance during the race.
2. Athletes who are confident that they will qualify for the next round (usually the top 3 boats) will conserve their energy at the end of the race and exhibit a Positive pacing profile.

References

[1] Chris R. Abbiss and Paul B. Laursen. Describing and understanding pacing strategies during athletic competition. *Sports medicine (Auckland, N.Z.)*, 38:239–52, 03 2008.

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

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

[4] Thomas Muehlbauer, Christian Schindler, and Alexander Widmer. Pacing pattern and performance during the 2008 olympic rowing regatta. *European Journal of Sport Science*, 10(5):291–296, 2010.

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

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

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 and the Carnegie Mellon Sports Analytics Club for hosting the reproducible research competition that pushed us to make our work reproducible and code and data 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. Finally, we would like to thank Hadley Wickham and those who have contributed to the development of the R packages in the **tidyverse** family. They are incredible contributions to the community and made everything in this analysis more efficient.