

# Pacing Profiles in 2000 Meter World Championship Rowing

*Dani Chu<sup>1</sup>, Ryan Sheehan<sup>1</sup>, Ming-Chang Tsai<sup>2,3</sup> and Jack Davis<sup>1</sup>* <sup>1</sup>*Department of Statistics and Actuarial Sciences, Simon Fraser University,* <sup>2</sup>*Canadian Sport Institute Pacific,* <sup>3</sup>*University of Toronto*

*2019-09-17*

## **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 predict a winning performance depends on the mode of sport. However, in 2000 meter 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 meter rowing races in a reproducible and 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](http://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. 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.

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

# 1 Introduction

Across “closed-loop” design sports, competitions where athlete(s) attempt to complete a set distance in the shortest time (Abbiss & 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; Michael D. Kennedy & Bell, 2003; Muehlbauer & Melges, 2011; Muehlbauer, Schindler, & Widmer, 2010). However, these attempts approach the problems in different ways and come to different conclusions. We attempt to standardize the definition of pacing profiles in rowing by using more granular data than other attempts.

## 1.1 Pacing Profiles in Rowing

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

In 2003 Kennedy and Bell 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 males, and an all-out profile was optimal for females.

In 2004, Garland used observational data from 2000 Olympic, 2001 World Championship and 2001 & 2002 British indoor Rowing Championship competitions. His analysis found that when using 4 time splits measured every 500 meters 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. in 2010 as well as Muehlbauer and Melges in 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 did affect pacing profiles in World Championship races between 2001 and 2009. They performed these analyses by fitting linear quadratic models to the four time splits.

## 1.2 Pacing Profiles in other Closed-Loop Sports

In other races of fixed distance like cycling, and running six pacing profiles have been defined (Abbiss & 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 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 meter sprint. This will often result in “negative-split” times in shorter events, and “positive-split” times in longer events.

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

Even pacing profiles are categorized by a relatively small portion of the race spend 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 shape 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 has a greater relative speed at the start of the race.

The last profile mentioned by Abbiss and Laursen is “Variable Pacing”. It is a strategy that is used to adapt to changing conditions in the race course, like uphill and downhill in cycling.

### 1.3 Our Approach

The classification of pacing profiles has usually been approached by fitting linear models to 4 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. We then hope to match the clusters we find to pacing profiles mentioned in racing literature. Furthermore, we then plan on determining which race factors affect the use of a pacing profile and which do not.

To do so we present the following:

1. A github repository with Global Positioning System (GPS), Media Start List, and Race Results data 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 summary of the race factors which did or did not have an association with pacing profiles.

## 2 Data Collection

The availability of easily usable data in rowing is limited. While some studies have data from ergometric machines (Michael D. Kennedy & Bell, 2003) and others have split time results (Garland, 2005; Muehlbauer & Melges, 2011; Muehlbauer et al., 2010), there has been limited use of the publicly available GPS race data. This is because the data is not stored in a local, easy to use area.

GPS data for almost every World Championship race from 2010 to 2017 is available in Portable Document Format (PDF) files that are located in the summary of each round of each category of each World Championship. Not only was it previously a long process to download each and every pdf file, but it was a long manual process to copy and paste the data from each file into a more convenient form. 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 5322 PDFs that were scraped from the website. Only 1 of these files was a broken link, which left us with 5321 PDF files from the 8 years of World Championships to use.

Table 1: Number of PDF Files by World Championship Year. The number of races fluctuates by year depending on whether it was an Olympic year, if Paralympic rowing was included and whether Junior World Championships were included.

Year	Number of PDF Files
2010	423
2011	839
2012	675
2013	681
2014	788
2015	1039
2016	201
2017	675

The next step to create our dataset was to extract the information from the PDF files. Some races were missing one or more of the GPS, Results, or Media Start List data. We did not parse any PDFs for races that were missing a file. This left us with 1,736 unique races over the 8 years of World Championships. We developed 3 separate PDF extractor functions and 1 main pdf extractor driver to parse each PDF and merge together information based on regular expressions and consistent locations of data within the PDFs. This process had to accommodate changes in naming conventions, changes to the location of information and addition of information that occurred over the years.

From the GPS PDF we extract average speeds and strokes per minute for each 50 metre split (sometimes smaller splits in later years) for each boat along with the race date, race number, round type, country abbreviation and lane number.

From the Results PDF we extract the finishing rank, the progression (next race or finishing tournament rank), 500 meter split times, whether the boat “Did not Start”, “Did not Finish” or was “Excluded” from the race, the country abbreviation and lane number.

From the Media Start List PDF we extract the names, birthdays and positions of each boat

member as well as the country abbreviation and lane number

We use parallel computing to speed up the process of extracting the information for the 5321 pdfs. Using 4 cores instead of one we see the time to extract files drop from 58 minutes to 24 minutes.

Finally we augment the data by breaking down the race’s category abbreviation. We can split each race’s category into Weight Class (Light or Open), Gender (Male, Female or Mixed), Size (1, 2, 4, 8), Discipline (Scull or Sweep), Adaptive (True, or False), Adaptive Designation (Arm and Shoulders, Intellectual Disability, Trunk and Arm, Leg Trunk and Arm, or None), Age Group (Junior or Senior), Race Round (Exhibition, Repechage, Heat, Quarterfinal, Semifinal, or Final) and a Qualifier or Final designation.

This left us with a dataset of 9264 boats’ races (rows) and 131 variables.

### 3 Data Filtering and Pre-Processing

Before beginning any clustering procedures we needed to clean the data set and keep only the boats’ races that we wanted to cluster. Some races have GPS data errors where the reported average speed is lower than the true average speed. In other cases the average speed is simply not reported. We remove any boat that has a unreported average speed at any of the split measurements (at every 50 meters) and any boat that saw reported average speed 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 (Garland, 2005). 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.

### 4 Clustering Pacing Profiles

With the boats’ races cleaned and speeds standardize to the same scale we can begin the clustering process. The idea is that we would like to group velocity curves of similar shape together. There is a large body of literature in clustering and the area of longitudinal clustering is growing. McNicholas and Subedi 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 (McNicholas, Sanjeena,

& Subedi, 2012). Additionally, Kumar and Futschik used a soft clustering technique to cluster the shapes of microarray data (Kumar & Futschik, 2007). 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 developed the k-Shape clustering technique for time series data (Paparrizos & Gravano, 2016).

## 4.1 k-Shape Clustering

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 our  $y_{i,j}$  defined above and transforms it to  $z_{i,j}$  defined as

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

A k-Shape algorithm therefore functions very similarly to the k-means algorithm (Lloyd, 1982) in the way that it uses an iteratively defines clusters to minimize within-cluster distance.

## 5 Clustering Results

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.

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

Cluster	Size	Average Distance Within
1	1951	0.1233
2	2227	0.0956
3	2548	0.1104
4	1444	0.162

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

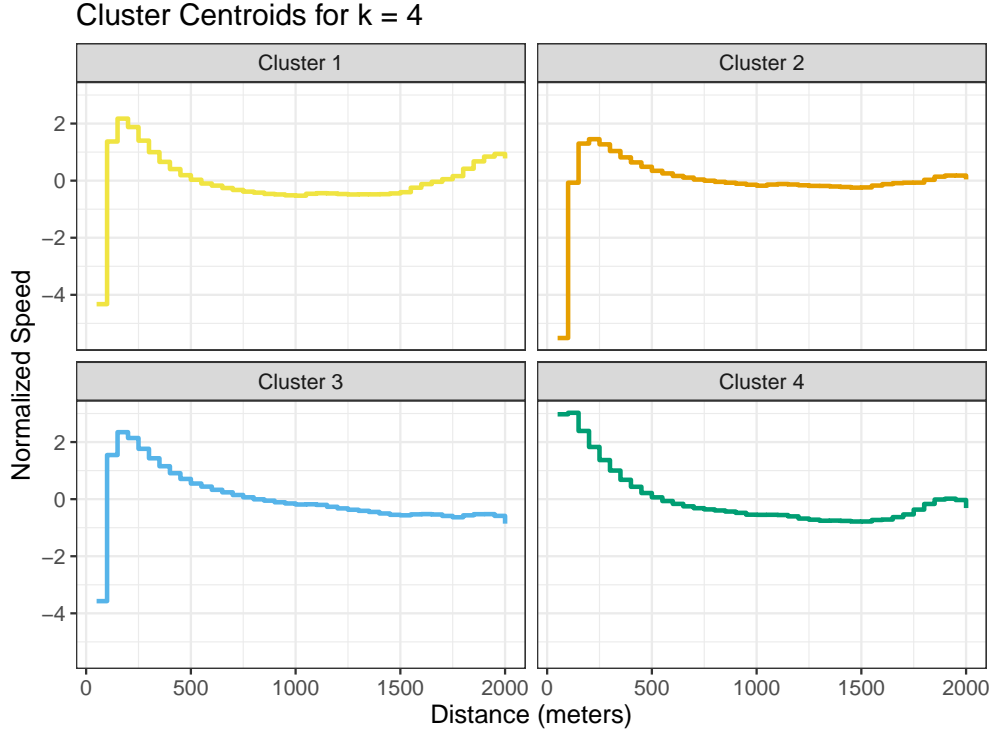


Figure 1: Cluster Centroids for k-Shape Clustering with 4 Clusters

the pacing curve without the effect of magnitude that size of boat, weight class and other variables would affect.

## 5.1 Identifying Pacing Profiles

We will now name the clusters based on the definitions given by Abbiss (Abbiss & 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 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.

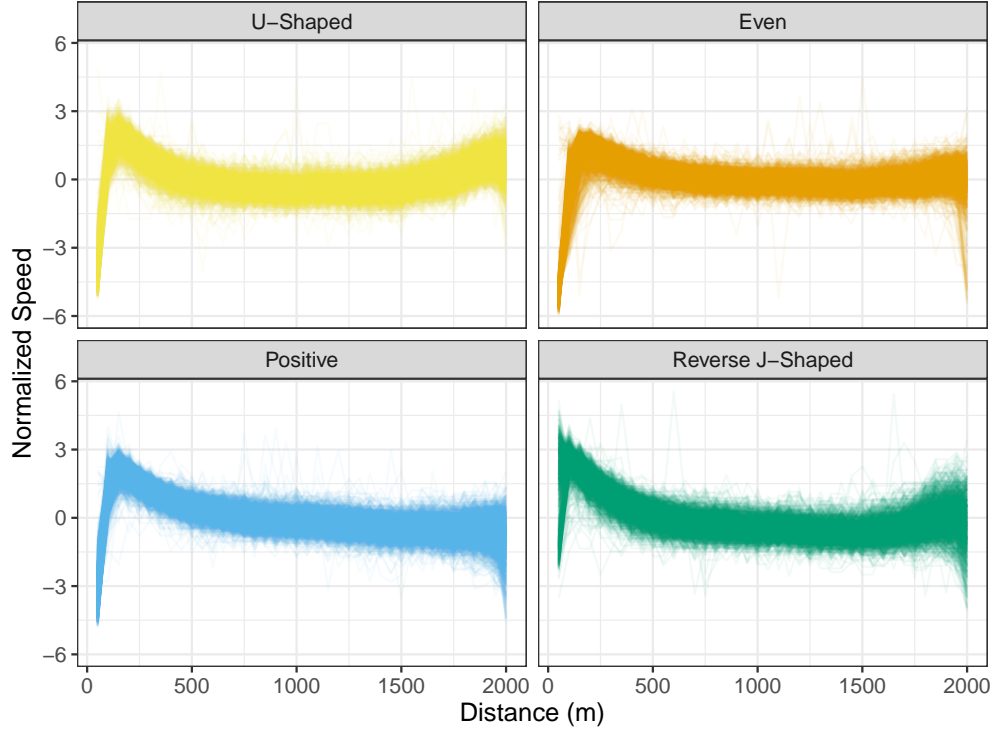


Figure 2: Pacing Profiles for k-Shape Clustering with 4 Clusters

## 5.2 Relationship between Boats Descriptors and Pacing Profiles

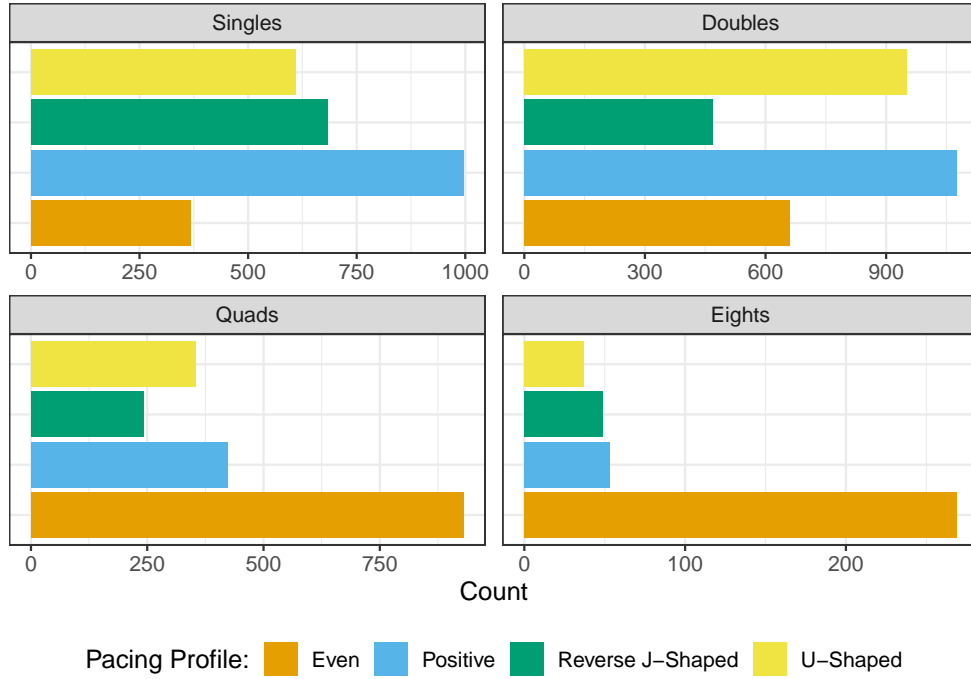
Now that we have identified and named the profiles. We will shift to examining what boats are more likely to exhibit a specific pacing profile. It is important to note that these comparisons will not indicate what is optimal, nor will it indicate what the rower or rowers intended before the race started.

### 5.2.1 Size of Boats

The first factor that we look into is the size of the boats in the race. 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, a 1-person boat does not. The 1-person boat enjoys more control to change pace at will. We can see that in the plot below that there does seem to be an effect.



Distribution of Pacing Profiles Given Boat Size



An 8-person boat is relatively more likely to exhibit a “Even” pacing profile than a “Positive” pacing profile. Next we describe the boat and race characteristics that affect the pacing profile that a boat exhibits.

## 6 Multinomial Logistic Regression

A multinomial logistic regression was fit 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 different 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 Bonferonni Correction applied so that our global Type II error rate is 0.05.

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	1.411
Size: Doubles	<b>0.4795</b>	<b>0.3802</b>	<b>0.6796</b>
Size: Quads	<b>0.1272</b>	<b>0.136</b>	<b>0.1607</b>
Size: Eights	<b>0.03556</b>	<b>0.0757</b>	<b>0.03537</b>
Heat or Final: Heat	<b>1.812</b>	1.037	<b>0.5397</b>

	Positive	Reverse J-Shaped	U-Shaped
<b>Race Placement: 2nd Place</b>	0.8631	1.02	1.207
<b>Race Placement: 3rd Place</b>	1.078	1.326	<b>1.498</b>
<b>Race Placement: 4th Place</b>	1.358	<b>1.604</b>	<b>1.6</b>
<b>Race Placement: 5th Place</b>	<b>1.76</b>	<b>1.928</b>	1.242
<b>Race Placement: 6th Place</b>	<b>3.162</b>	<b>3.16</b>	1.205
<b>Discipline: Sweep</b>	<b>1.814</b>	1.201	<b>1.966</b>
<b>Gender: Women</b>	<b>1.883</b>	<b>1.683</b>	<b>1.663</b>
<b>Weight Class: Open</b>	<b>1.432</b>	<b>1.523</b>	<b>1.284</b>

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 column for size 1 boats. This is because the size 1 boats are the baseline for the boat size variable.

So if we first look at the value for the “Positive” pacing profile for the size 8 boats. The odds that a size 8 boat would follow a “Positive” pacing profile is 0.03 times as large as the odds that a 1 person boat would follow an “Even” pacing profile holding all other variables constant. We can see that since all odds ratios for the size 8 boats are less than 1 that an 8 person boat is most likely to exhibit an “Even” pacing profile. This aligns with what we saw in the graph.

### 6.0.1 Round of Race (Heat or Final)

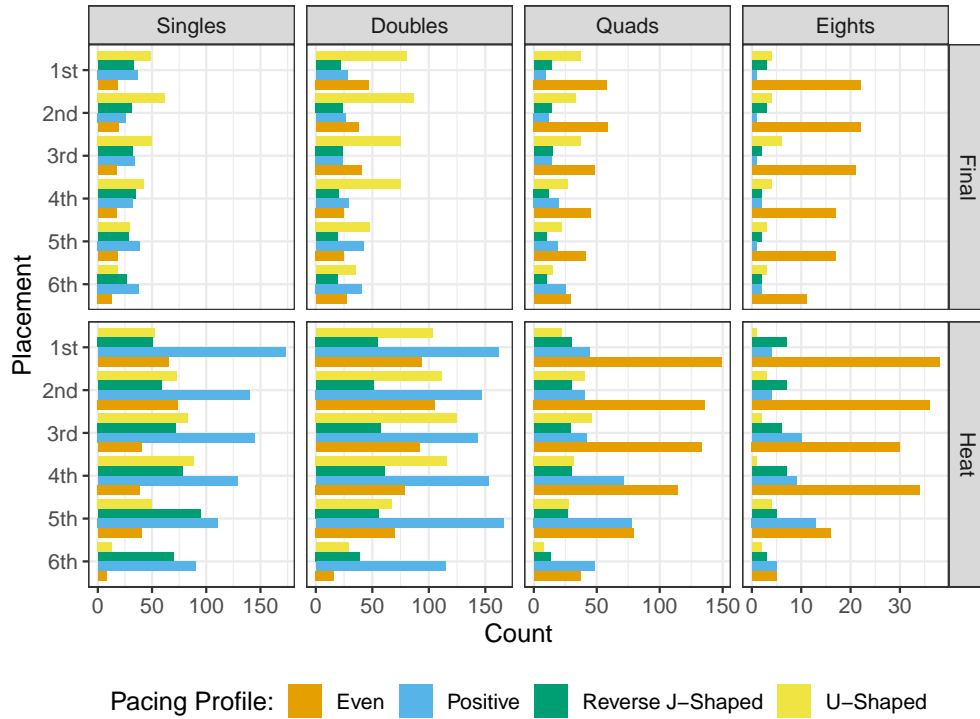
Next we can see that holding all other variables constant that the “Positive” pacing profile is nearly 2 times more likely to be used in a heat then a final. This would make sense as boats that are in heats are more likely to want to conserve their energy for their future races. Interestingly, in rowing there is a rule that disqualifies boats that are not “trying their hardest” throughout the race (NEED CITATION). This shows that boats are using a pacing profile with positive split times in heats but this could be intentional or unitentional.



This plot that accounts only for size and round, shows that boats are more likely to follow “Positive” pacing profiles in Heats and to use “U-Shaped” pacing profiles in finals. In bigger boats it is clear that the “Even” pacing profile is still the most used

## 6.0.2 Race Placement

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’s is whether this would affect how the boat 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 to follow a given pacing profile. One explanation could be that in most races, the first 3 boats are the ones to qualify for the next race. It would therefore make sense if they followed a similar distribution of pacing profiles. Unfortunately looking for an interaction between these two variables would require more data than we have available and so we leave this for further investigation.

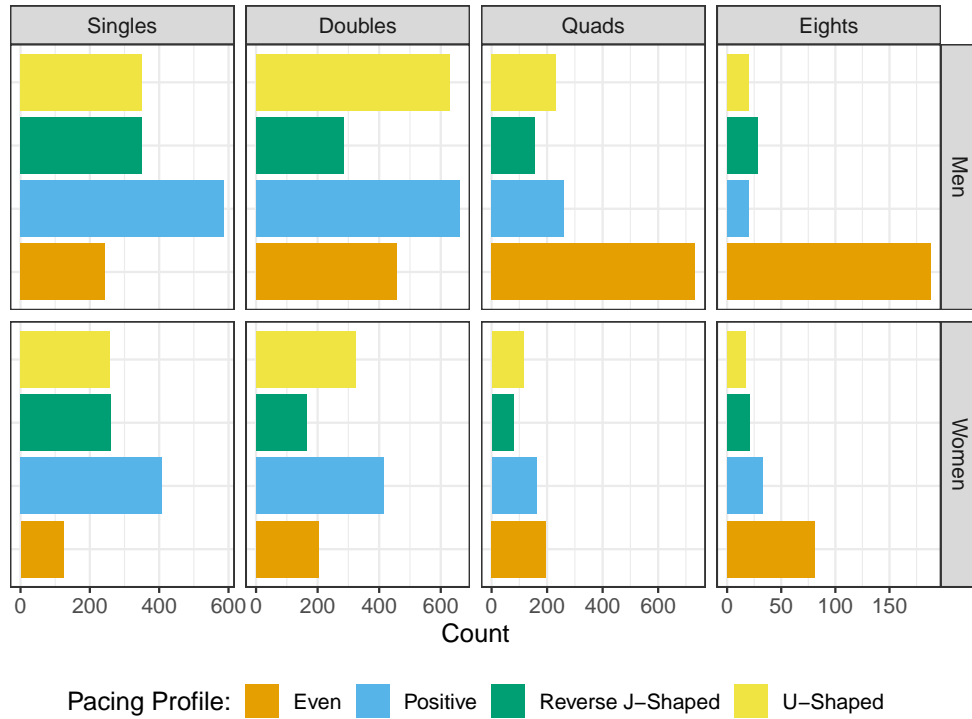


### 6.0.3 Discipline

A boat's discipline is either called a "Stroke" or a "Sweep". "Stroke" 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 "Stroke" boats.

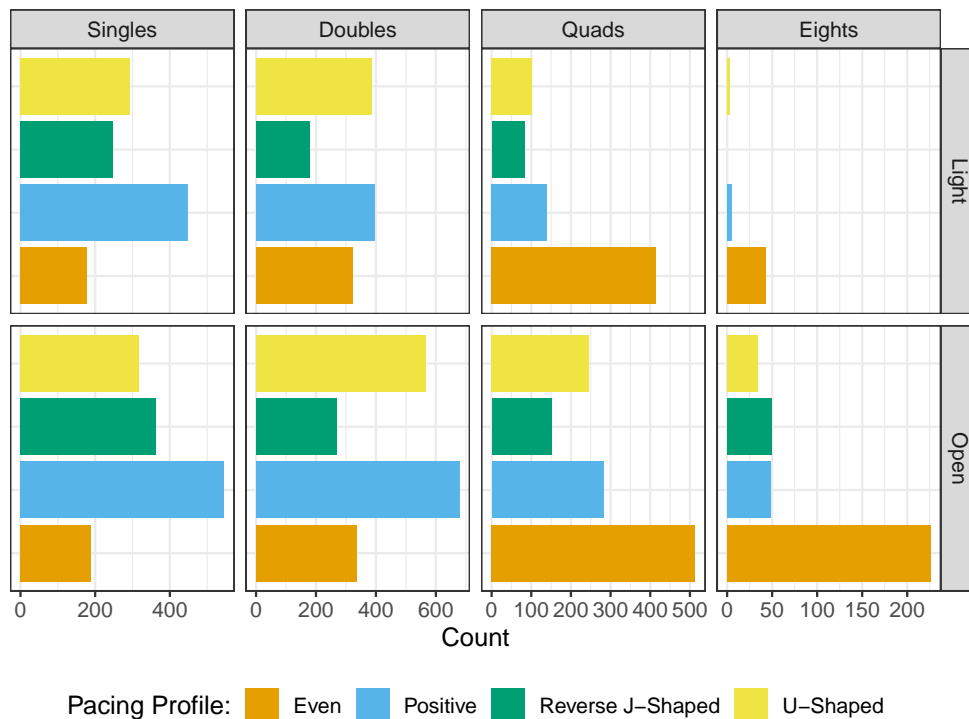
### 6.0.4 Gender

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



### 6.0.5 Weight Class

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.



## 7 Conclusions

We have identified pacing profiles in a more data-driven way than previous work.

We found which variables affected the likelihood of following a given pacing profile and by how much.

## 8 Future Research

Investigate optimal pacing profiles using more testing now that pacing profiles have been identified.

Investigate progression of pacing profiles for boats throughout an event.

### 8.1 Data Accessibility

The other contribution of this work is not only the analysis and identification of these observed pacing profiles but the data that is now accessible for use in further research by the open community. We welcome other researchers to not only check our work but further the field of analytics in rowing by using the rich dataset provided at [https://github.com/danichusfu/rowing\\_pacing\\_profiles](https://github.com/danichusfu/rowing_pacing_profiles). Not only is the data available in csv format, but the code to scrape the pdfs from [www.worldrowing.com](http://www.worldrowing.com), and extract the data from the pdfs. This way the data is available for use and our work can be improved upon by collecting more data of interest from the pdfs.

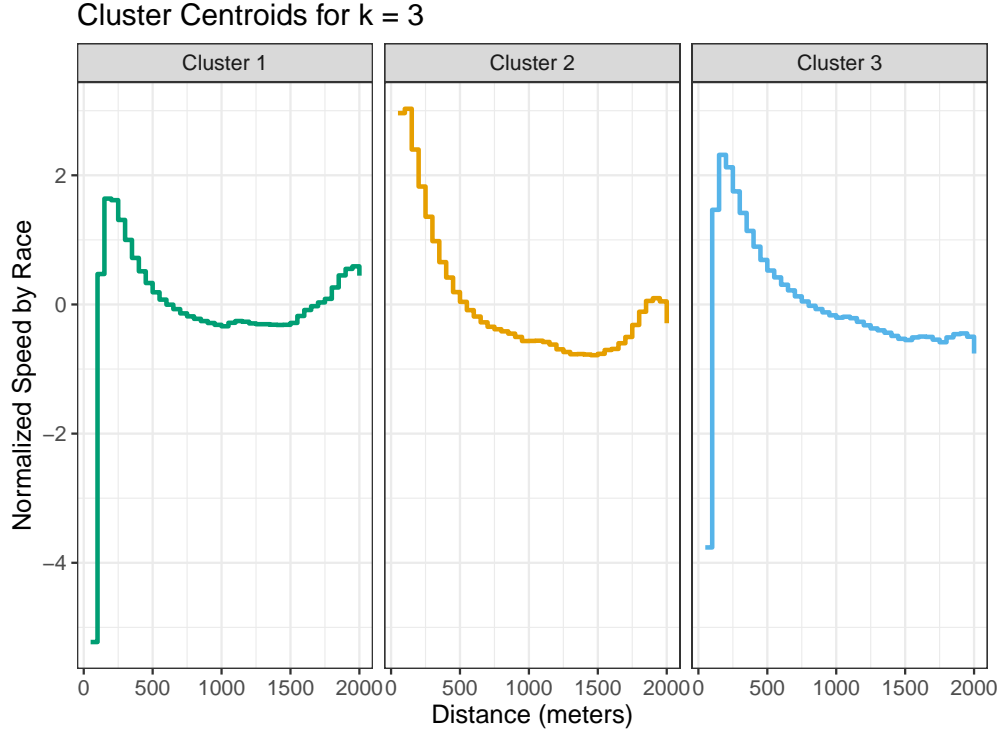


Figure 3: Cluster Centroids for k-Shape Clustering with 3 Clusters

## 9 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 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.

## 10 Appendix

Note, all plots use a colourblind friendly palette

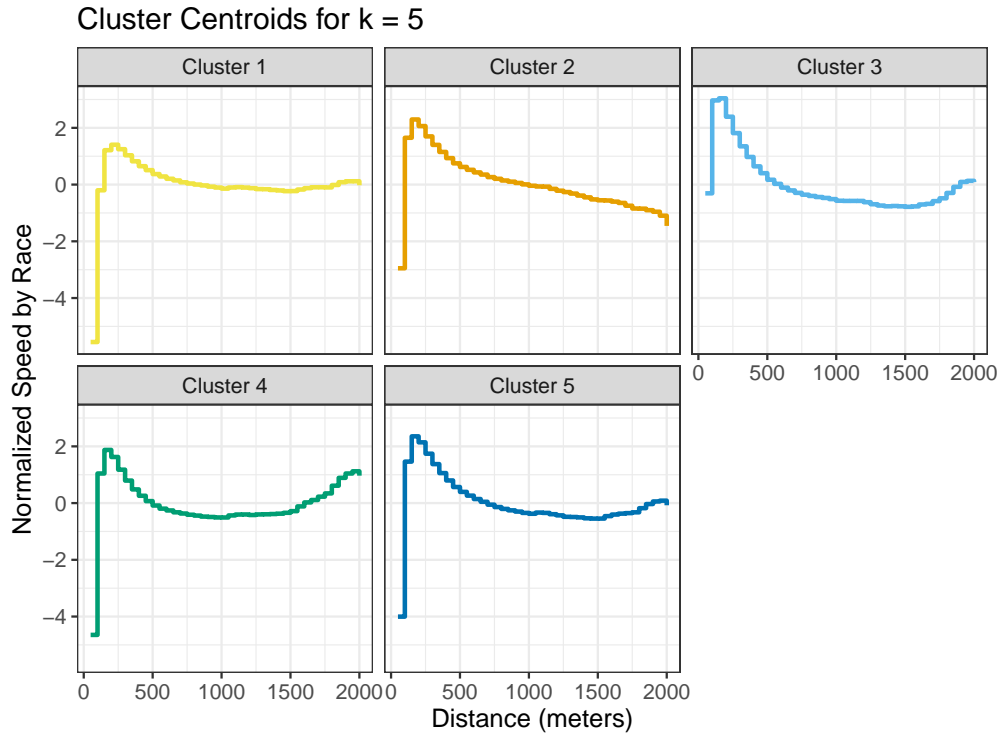


Figure 4: Cluster Centroids for k-Shape Clustering with 5 Clusters

## References

- Abbiss, C. R., & Laursen, P. B. (2008). Describing and understanding pacing strategies during athletic competition, *38*, 239–52.
- Betebenner, D. W. (2017). *toOrdinal: Function for converting cardinal to ordinal numbers by adding a language specific ordinal indicator to the number*. Retrieved from <https://centerforassessment.github.io/toOrdinal>
- Chen, Y., Keogh, E., Hu, B., Begum, N., Bagnall, A., Mueen, A., & Batista, G. (2015). *The ucr time series classification archive*. Retrieved from [www.cs.ucr.edu/~eamonn/time\\_series\\_data/](http://www.cs.ucr.edu/~eamonn/time_series_data/)
- Csárdi, G., Wickham, H., Chang, W., Hester, J., Morgan, M., & Tenenbaum, D. (2017). *Remotes: R package installation from remote repositories, including 'github'*. Retrieved from <https://CRAN.R-project.org/package=remotes>
- Daróczi, G., & Tsegelskyi, R. (2018). *Pander: An r 'pandoc' writer*. Retrieved from <https://CRAN.R-project.org/package=pander>
- 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. British Association of Sport; Exercise Medicine. Retrieved from <https://bjsm.bmj.com/content/39/1/39>
- Grolemund, G., & Wickham, H. (2011). Dates and times made easy with lubridate. *Journal*



- of *Statistical Software*, 40(3), 1–25. Retrieved from <http://www.jstatsoft.org/v40/i03/>
- Henry, L., & Wickham, H. (2018). *Purrr: Functional programming tools*. Retrieved from <https://CRAN.R-project.org/package=purrr>
- Hester, J. (2018). *Glue: Interpreted string literals*. Retrieved from <https://CRAN.R-project.org/package=glue>
- Kumar, L., & Futschik, M. (2007). Kumar l, futschik e.. mfuzz: A software package for soft clustering of microarray data. *Bioinformation* 2: 5-7, 2, 5–7.
- Leeper, T. J. (2018). *Tabulizer: Bindings for tabula pdf table extractor library*.
- Lloyd, S. P. (1982). Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28, 129–137.
- McNicholas, P., Sanjeena, & Subedi. (2012). Clustering gene expression time course data using mixtures of multivariate t-distributions, 142, 1114–1127.
- Michael D. Kennedy, & 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.
- Muehlbauer, T., & Melges, T. (2011). Pacing patterns in competitive rowing adopted in different race categories. *The Journal of Strength & Conditioning Research*, 25. Retrieved from [https://journals.lww.com/nsca-jscr/Fulltext/2011/05000/Pacing\\_Patterns\\_in\\_Competitive\\_Rowing\\_Adopted\\_in.15.aspx](https://journals.lww.com/nsca-jscr/Fulltext/2011/05000/Pacing_Patterns_in_Competitive_Rowing_Adopted_in.15.aspx)
- Muehlbauer, T., Schindler, C., & Widmer, A. (2010). Pacing pattern and performance during the 2008 olympic rowing regatta. *European Journal of Sport Science*, 10(5), 291–296. Routledge. Retrieved from <https://doi.org/10.1080/17461390903426659>
- Ooms, J. (2018). *Pdftools: Text extraction, rendering and converting of pdf documents*. Retrieved from <https://CRAN.R-project.org/package=pdfutils>
- Paparrizos, J., & Gravano, L. (2016). K-shape: Efficient and accurate clustering of time series. *SIGMOD Rec.*, 45(1), 69–76. New York, NY, USA: ACM. Retrieved from <http://doi.acm.org/10.1145/2949741.2949758>
- R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <https://www.R-project.org/>
- Sarda-Espinosa, A. (2018). *Dtwclust: Time series clustering along with optimizations for the dynamic time warping distance*. Retrieved from <https://CRAN.R-project.org/package=dtwclust>
- Urbanek, S. (2018). *RJava: Low-level r to java interface*. Retrieved from <https://CRAN.R-project.org/package=rJava>
- Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from <http://ggplot2.org>
- Wickham, H. (2017a). *Tidyverse: Easily install and load the 'tidyverse'*. Retrieved from <https://CRAN.R-project.org/package=tidyverse>

- Wickham, H. (2017b). *Multidplyr: Partitioned data frames for 'dplyr'*. Retrieved from <https://github.com/hadley/multidplyr>
- Wickham, H. (2018a). *Stringr: Simple, consistent wrappers for common string operations*. Retrieved from <https://CRAN.R-project.org/package=stringr>
- Wickham, H. (2018b). *Forcats: Tools for working with categorical variables (factors)*. Retrieved from <https://CRAN.R-project.org/package=forcats>
- Wickham, H., François, R., Henry, L., & Müller, K. (2018a). *Dplyr: A grammar of data manipulation*. Retrieved from <https://CRAN.R-project.org/package=dplyr>
- Wickham, H., Hester, J., & Chang, W. (2018b). *Devtools: Tools to make developing r packages easier*. Retrieved from <https://CRAN.R-project.org/package=devtools>
- Xie, Y. (2014). Knitr: A comprehensive tool for reproducible research in R. In V. Stodden, F. Leisch, & R. D. Peng (Eds.), *Implementing reproducible computational research*. Chapman; Hall/CRC. Retrieved from <http://www.crcpress.com/product/isbn/9781466561595>
- Xie, Y. (n.d.). *Tinytex: Helper functions to install and maintain 'tex live', and compile 'latex' documents*. Retrieved from <https://github.com/yihui/tinytex>