Pacing Profiles in 2000 Meter World Championship Rowing

2018-09-16

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 2000m 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 2000m rowing races. To do this the average speed and stroke rate (SR) for each 50m split for each boat in every race of the Rowing World Championships from 2010-2017 was Scraped from www.worldrowing.com. Pacing profiles are determined by using k-Shape clustering on the average boat speeds at each 50m split. Finally, clusters are described using boat and race descriptors to draw conclusions about who, when and why each pacing profile was observed. Chi-Squared Tests of Independence with Bonferroni corrections are used 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) and rank are all found to affect pacing profiles. Whereas, gender, and weight class do not affect a boat's pacing profile. This novel approach of using clustering is able to objectively define four strategies used in 2000m rowing competitions.

Keywords: rowing, pacing profiles, k-Shape clustering, 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 differnt 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 (Muehlbauer, Schindler, & Widmer, 2010) (Michael D. Kennedy & Bell, 2003) (Muehlbauer & Melges, 2011) (Garland, 2005). However, these attempts approach the problems in different ways and come to different conlclusions. We attempt to standardize the definition of pacing profiles in rowing by using more granual 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 et al., 2010) (Muehlbauer & Melges, 2011).

In 2003 Kennedy and Bell used simulated rowing and training results to suggest that there was a 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 Meuhlbauer et. al. in 2010 and Meuhlbauer 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 sculle did not affect pacing strateiges for the 2008 Olympics. In 2011, they had a different finding that indicated that round of race did effect 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 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 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 maximal 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 2 sub pacing 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 fo 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 uphills and downhills 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 Sytem (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 the current data.
2. A novel approach of classifying pacing profiles for boats in 200m rowing.
3. Which race factors affect and do not affect the use of a given pacing profile during an event.

2 Data Collection

The availability of 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 et al., 2010) (Muehlbauer & Melges, 2011), there was 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 Fomat (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 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 roguhly 25 minutes to run. In total there were 5322 PDFs that were scraped from the website. 1 of these files was a broken link so it left us with 5321 PDF files from the 8 years of World Championships.

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 developped 3 separate PDF extractor functions and 1 main pdf extractor driver to parse each PDF and merge together information based on regular expressions and consitent locations of data within the PDFs. This process had to accomodate changes in naming conventions, changes to the location of information and addition of information that occured over the years.

From the GPS PDF we extract average speeds and strokes per minute for each 50 meter split (sometimes smaller splits in later years) for each boat along with the race date, race nuber, 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 abbreivation 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 cateogry 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, Intelectual Disabitliy, Trunk and Arm, Leg Trunk and Arm, or None), Age Group (Junior or Senior), Race Round (Exhibition, Repecharge, 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' racesthat 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 with errors removed 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 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 developped 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. When SBD is evaluated against other distance metrics such as Dynamic Time Warping, it reaches similar error rates on the UCR datasets but much more efficient run times.

The k-Shape algorithm is implemented in the dtwclust package (Sarda-Espinosa, 2018). In it's 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 below.

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

k-Shape then functions very similarly to the k-means algorithm (Lloyd, 1982) in the way that it uses an interative refinement that minimizes a given distance function.

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, there is an iteration where cluster memberships do not change, for our given seed. The size of the clusters and average distance are reported in Table 2.

Table 2: Summary of clusters for k = 4

Cluster	Size	Average Distance
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, that is to be expected as they are all races, there are distinct features that separate them. The centroids are plotted with respected 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.

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 accelartion to a moderate peak velocity, a slow middle section and a final sprint that almost reaches peak velocity. This coordiantes 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 acceleartion 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.

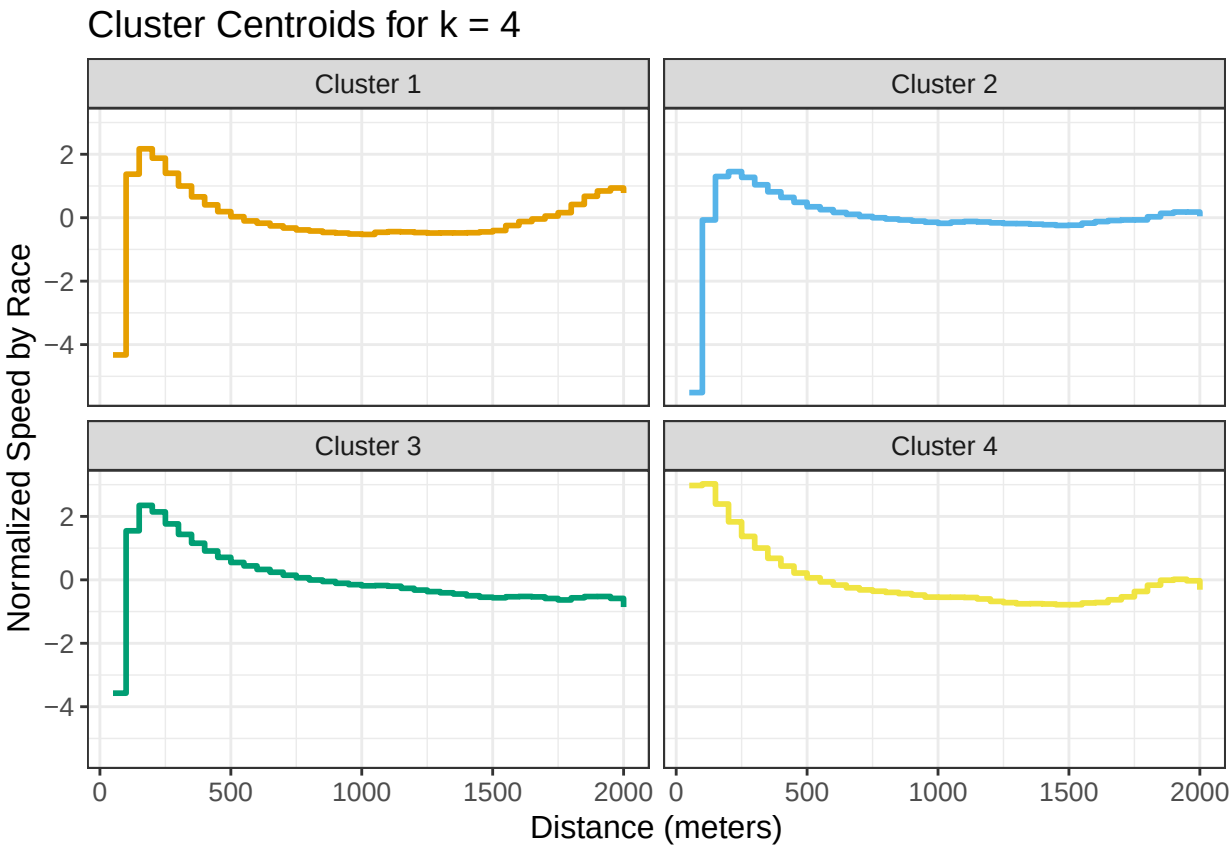

Figure 1: Cluster Centroids for k-Shape Clustering with 4 Clusters

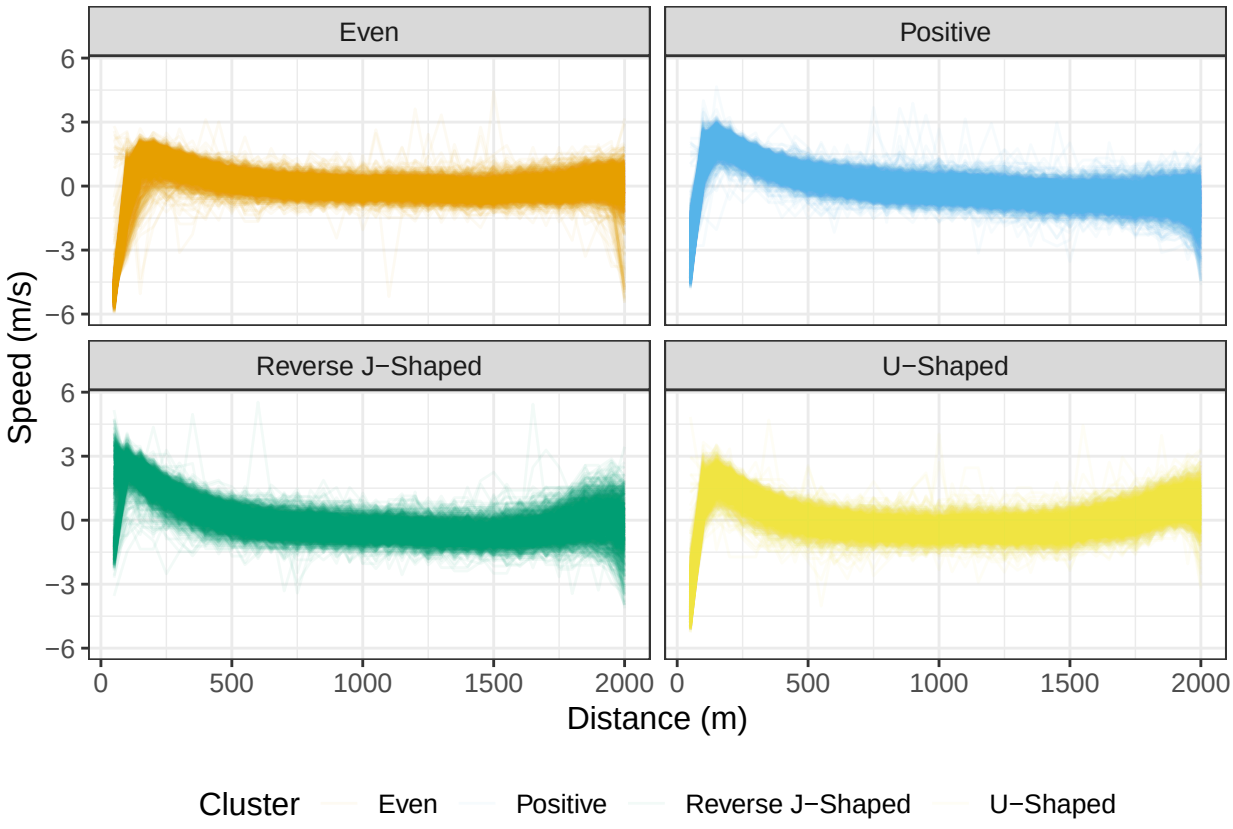

Figure 2: Pacing Profiles for k-Shape Clustering with 4 Clusters

5.2 Relationship between Boats Descriptors and Pacing Profiles

Now that we have identified and named the profiles

5.2.1 Size of Boats

5.2.2 Round of Race

5.2.3 Final Placement

5.2.4 1st Half Rankings

5.2.5 Gender

5.2.6 Weight Class

5.2.7 Stroke Rate

5.2.8 Adaptive Races

6 Conclusions

7 Future Research

7.1 Data Accessibility

7.2 Extending the Methods

8 Acknowledgements

9 Appendix

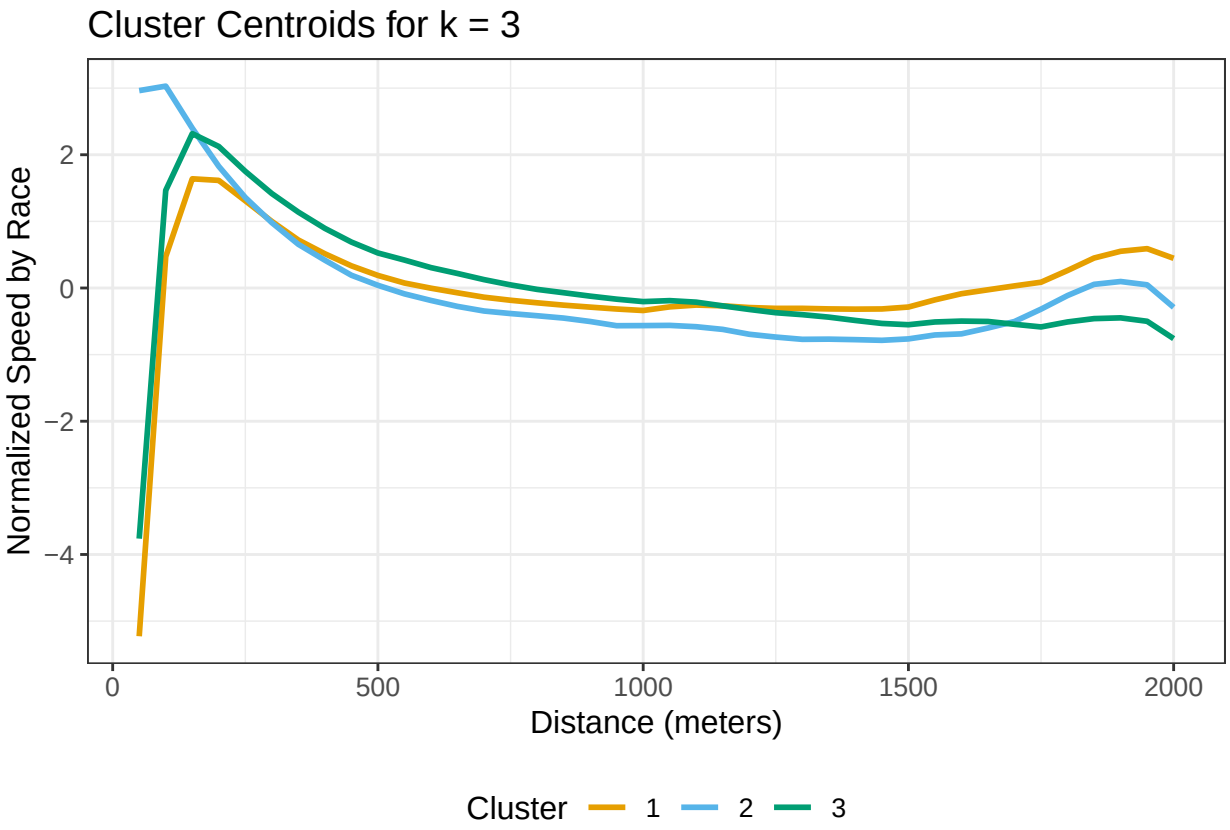

Figure 3: Cluster Centroids for k-Shape Clustering with 3 Clusters

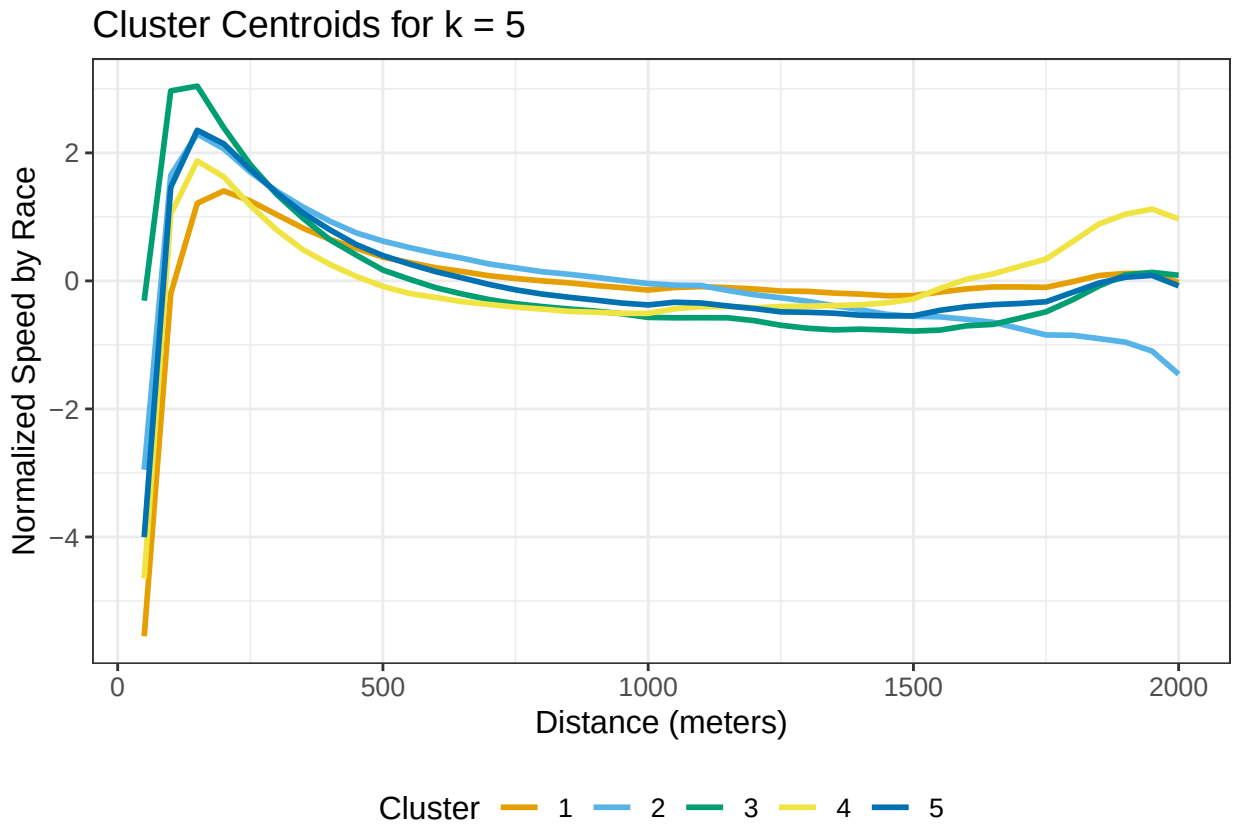

Figure 4: Cluster Centroids for k-Shape Clustering with 5 Clusters

References

Abbiss, C. R., & Laursen, P. B. (2008). Describing and understanding pacing strategies during athletic competition, *38*, 239–52.

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/

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

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; Excercise 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/

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

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=pdftools

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., Hester, J., & Chang, W. (2018). *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