Matthew Mulligan
Molly Graber
Chris Hitte
Jonathan Magiera
William Diment

Distance Runner Classification

Background:

Distance running has played a large role throughout much of human history. Recently, it has been proposed that humans evolved as long distance runners to compete with other carnivorous predators. From the time of the first Olympics, distance running has in turn become a spectacle to observe and compete in. For much of human history man has asked, "how can I get faster?" More recently, the collection of large scale data for training runs with GPS watches has made it possible to test potential models for distance running.

We will attempt to develop a novel approach to modeling optimal training paces for distance runners. Physiologically, we note that to run different distances, the human body must undergo different processes in order to achieve optimal performance levels, namely exercising slightly anaerobically for relatively shorter distances, and purely aerobically for more extended distances. There appears to be a physiological split, which occurs at a *crossover point,* where an individual is able to maintain his or her maximal oxygen consumption. Note, that while it is possible to exceed this level of intensity, the human body may only maintain such *anaerobic* processes for a short period of time. Similarly, the body may maintain a slightly slower *aerobic* pace for a much more extended period of time.

Though many exercise physiologists have proposed different training models (Daniels, Gilbert, etc.), the majority of the approaches are empirically derived, and generally overfit smaller datasets collected by individual coaches. Similarly, though there exist platforms which attempt to capture large datasets (Garmin, Strava, etc.), often the variability in data across different terrains may only be approximated, and thus, fails to capture true training paces.

Proposed Objective:

We will attempt to examine the performance of collegiate distance runners at varying distances in track races in order to better quantify the physiological split across middle and longer distances. We choose collegiate athletes due to the large availability of data sets, low variability relative to younger athletes who may see large discrepancies from athlete-to-athlete based on maturity, and wide range of race performances. Moreover, we choose to examine exclusively track races to minimize variability that may not be captured over cross country or road races. Though there may exist slight fluctuations by region in track and field races, this factor will be minimized relative to races which may not take into account factors such as weather conditions, differences in course, or elevation. Though this set may be somewhat restrictive in the sense that it biases toward fitter athletes, we hope to capture performance of collegiate athletes across all levels to develop the best model possible.

Project Proposal:

In order to assess the validity of our objective, we will first develop and employ a data scraper on TFRRS.org, a web database which contains race results over the last ten years across NCAA Division I, II, and III, NAIA, and NJCAA. Though we will need to be careful in examining the data we collect, we will focus on examining athletes with results for multiple races distances both above and below the predicted physiological split. Traditionally, it has been shown that most fit athletes appear to be able to maintain their maximal oxygen consumption for 6-8 minutes.

Using the data we gather, we hope to use a K-means clustering approach to approximate race performances. Because athletes traditionally appear to favor either the middle or long distance regime rather than performing equally well in each, choosing a K-means over a traditional logistic regression approach may be the optimal choice to help find similar points rather than approximate overall physiology. Based on these results, we hope to be able to choose only a single or pair of race times in order to predict runner performance across all distances.