Cross Country Course Standardization

Arianna DeBoer and Annie Wicker

November 2024

1 Project Description, Background, and Motivation

Cross country courses can vary significantly in difficulty based on factors like terrain (flat or hilly), surface (grass or dirt), weather, and altitude. Comparing runners' times between different courses can lead to misleading conclusions about capabilities and future race results. While it is impossible to predict how runners will do on a given course and day, it is possible to retrospectively quantify how hard a given race was based on how runners performed compared to past performances. The Track and Field Results Reporting System (TFRRS) is a reputable platform that publishes race results for cross country races across the country. However, it does not have a way to compare results from different courses.

The goal of this project is to create an automated workflow that collects data from TFRRS to standardize runners' times across cross country courses. The goal is to be able to compare race times on courses that may have differed in terrain, elevation, and other factors. Though our analysis will not directly measure the effect of weather fluctuations or specific terrains, it will compare runners' times across different courses and different days to predict how they will do on a future course. This will prove a useful tool for collegiate coaches to allow for more accurate comparison of runners' performances.

2 Data Description

The Track and Field Results Reporting System (TFRRS) is a platform that compiles data from college cross country and track and field competitions. This website obtains data from individual timing companies and uniformly organizes the results. Each meet has its own webpage which contains results from every race run at that meet. For example, the page for 2024 Panorama Farms Invitational has results from both the women's six kilometer race and the men's eight kilometer race. The sample page described can be found through this link.

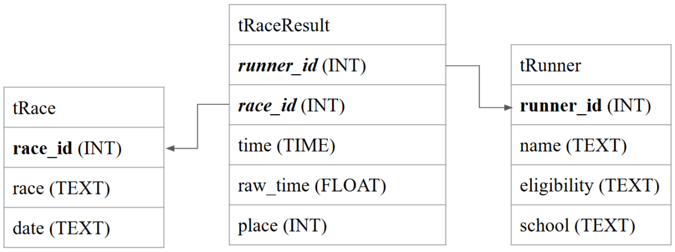

Figure 1: Database Entity Relationship Diagram

Each race result page on TFRRS contains the name of the race, the date it was run, and the distance of both the men's and women's courses. The page also contains information about each person who ran in the race.

Our workflow scrapes TFRRS data given a race results page URL. A user can specify that they want to scrape men's or women's results, and the scraped data is compiled into pandas dataframes. The dataframe contains the data described above: the place each runner finished (in that specific race), their name (First Last), year, college, average mile pace, total race time, score (the runner's place as a point value), the name of the meet, and the date this meet was run. These data are recorded for every runner who ran the race whose URL is entered.

Scraped data is loaded into a relational database. All column names are the same as in a TFRRS results table, with the exception of ”time” and ”raw time”. The raw time column formats the time column to be in MM:SS.MS, while the time column lists the time in total seconds. The data is stored in a sqlite database, and the accompanying tables are shown in the below entity relationship diagram.

3 Methods

The compare_two_courses() function calculates the average differences in time between two courses. It has the following workflow and mathematics:

- Two courses are entered by their ID's. The first is called the primary course and is used as a reference point. The second is called the secondary course.
- All runners who competed in both of the specified races are located. These are the only times used to compare the two races in order to prevent

selection bias. The number of runners used in the comparison is outputed as 'NumCompared'.

- The times from these runners are averaged for the primary course as well as the secondary course. This number is then saved.
- The average difference in seconds is found by subtracting the average time for the secondary course from the average time for the primary course. This is outputted as 'Difference'.
- A positive value for 'Difference' means that the secondary course was slower than the primary course.
- The ratio of the two courses is found by dividing the secondary course's average time by the first course's. This is outputted as 'Ratio'.
- A value of 'Ratio' that is greater than 1 means that the secondary course was slower than the primary course.

The conversions() function calculates how fast each loaded race was in comparison to an inputted primary course. These calculations are made both as time differences and as ratios. The function has the following workflow and mathematics:

- One course is entered by it's ID. This course is set as the primary course
- A minimum number of comparisons is entered as an integer, set to min_comparisons = 15 if not specified. There must be at least that many runners in common for a race to be converted. For clarity, for the rest of the explanation, we refer to this number as 15.
- First, all races with at least 15 runners in common with the primary course are designated secondary courses. The compare_two_courses() function is run with the primary course and each secondary course. Results for ratio and time difference are entered into a dataframe as ratio_conversion and time_conversion.
- Next, all non-secondary races with at least 15 runners in common with either the primary or secondary courses are designated tertiary courses. The compare_two_courses() function is run with each tertiary course and each primary or secondary course it shares runners with. In order to find the ratio conversion for each tertiary course, the results from each course comparison are weighted based on how many runners were compared. The calculation for each looks something like this:

  tertiary ratio $= 4/15 * (1 \rightarrow 2) * (2 \rightarrow 4) + 9/15 * (1 \rightarrow 3) * (3 \rightarrow 4) + 2/15 * (1 \rightarrow 4)$

  where the ratio of course 1 to course 4 is being calculated, $(1 \rightarrow 2)$ is the ratio conversion from course 1 to course 2, and 4 runners are in common between these two courses out of 15 total comparisons available for course 4. A similar calculation is completed for the time difference in seconds:

tertiary difference $= 4/15 * ((1 \rightarrow 2) + (2 \rightarrow 4)) + 9/15 * ((1 \rightarrow 3) + (3 \rightarrow 4)) + 2/15 * (1 \rightarrow 4)$

where the difference in seconds from course 1 to course 4 is being calculated, $(1 \rightarrow 2)$ is the time conversion from course 1 to course 2, and 4 runners are in common between these two courses out of 15 total comparisons available for course 4.

- This same process is repeated for all other courses that have not yet been converted. These courses are called quaternary courses if they share at least 15 runners in common with any combination of primary, secondary, or tertiary courses.
- At this point, any courses that have not been classified as primary, secondary, tertiary, or quaternary are left out the analysis. The user is alerted of this. All other results are returned in a dataframe.

The virtual_race() function takes a list of schools and runs a virtual race between them based on the times of runners in these schools. The function has the following workflow:

- The conversions() function is called to calculate standardized ratios for all meets.
- These conversions are applied to all runners from the inputted schools.
- The converted times are averaged for each runner.
- The runners are ordered by average time.

4 Limitations

The compare_two_courses() function course works best when there are many points of comparison. Courses with few runners in common will have less meaningful results. Additionally, all runners in common must be running the course as a race for best results. If a team ran one meet as a workout, for instance, a comparison using this race will be less accurate. This function also assumes that runners have relatively little gain in fitness over the course of the season. If an entire team became much faster between two meets the comparison would be less accurate.

The conversions() function works best when the primary course has many points of comparison with other courses. Because of this, the primary course should be chosen carefully, and the function should be re-run with a different primary course if not enough courses are able to be compared.

The virtual_race() function works best when comparing teams who race in meets that can be connected. Additionally, if at any point one of the teams purposefully runs a meet without giving their full effort, for instance, as a workout, the results will be less accurate.

5 Future Goals

As we continue this project, we plan to create a function that chooses which course should be inputted as the primary course in the conversions() function in order to maximize connections and therefore accuracy. We also plan to create a function that scores the virtual meet and lets the user know which team won.

6 Group Work

Annie wrote the webscraping script to get the data from TFRRS. Arianna created the database and loaded the scraped data into it. Both of us wrote functions to query the database, and Arianna wrote the major function to map courses by relative difficulty. Annie created the dashboard and incorporated the database functionality into it.