DataFest 2019 @ OSU

Information and Documentation

April 6, 2019

Contents

Welcome to DataFest 2019 @ The Ohio State University! This document will serve as your guide to the weekend. For a schedule of events, see here:

- https://data-analytics.osu.edu/datafest/events

Please be aware of the many other files that you have:

- `asa-conduct-policy.pdf`
  - Information about the ASA's conduct policy which is in place for this event.
- `datafest-2019-codebook.xlsx`
  - Documentation for this year's dataset!
- `datafest-2019-data-policy.pdf`
  - Information about using this year's data.
- `datafest-2019-evaluation-sheet.pdf`
  - Information about how judges will evaluate presentations.
- `datafest-2019-presentation-guidelines.pdf`
  - Information about expectations and rules for presentations.
- `datafest-2019-q&a.pdf`
  - Questions and answers from already completed DataFests.
- `datafest-2019-student-guidelines.pdf`
  - General logistical information about the event.
- `data/`
  - The folder which contains the four `.csv` files which are this year's data!!!

The Challenge

The Canadian National Women's Rugby Team seeks your advice on the role of workload and fatigue in Rugby 7's. Rugby 7's is a fast-paced, physically demanding sport that pushes the limits of athlete speed, endurance and toughness. Rugby 7's players may play in up to three games in a day, resulting in a tremendous amount of athletic exertion. Substantial exertion results in fatigue, which may lead to physiological deficits (e.g., dehydration), reduced athletic performance, and greater risk of injury.

Despite the importance of managing player fatigue in professional athletics, very little is known about its effects, and many training decisions are based on "gut feel." Currently, training load is measured through a

combination of subjective measurements (asking players how hard they worked) and objective measurements from wearable technology. Fatigue is typically estimated by asking players how they feel in wellness surveys. However, there is no agreed-upon standard of defining fatigue so the relationship between workload and fatigue is unclear. In this challenge, we encourage you to explore new ways of measuring fatigue and examine its effects on players' performance and physical wellness. The datasets provide a number of observations that we believe will be useful to measure fatigue in players of the Canadian National Women's Rugby Team in the 2017 - 2018 season including the Rugy World Cup Sevens. Remember that training load is not the same as fatigue, and one question to explore is whether you can find evidence that some measures of training load are better than others.

Some issues to consider:

1. How reliable are subjective wellness data? Can you quantify the individual variation in self-reported data and use this to adjust measures of wellness?
2. Should the quality of the opponent or the outcome of the game be considered when examining fatigue during a game?
3. Some accepted (and even widely used) measurements of training load or fatigue are naive. For example, you'll find in these data a "Monitoring Score" which simply sums the values of other subjective scores in an attempt to create a single overall measure of fatigue. Is a simple sum useful? Or can it be improved? For example, are all components of this Monitoring Score needed? Are some more important than others, and why?
4. Be wary of missing observations. Most often they indicate that a player simply did not provide information or that sensors were not functioning. But in some situations values are missing because they are not meaningful in a certain context. You'll find that a one-size-fits-all approach is not useful.
5. You will find it tempting to use the location data to help inform on-field strategy. **We advise against this because it is unlikely to help you understand fatigue.** The location data are provided in order to help you study fatigue. For example, it could be used to, verify hypotheses, or evaluate player fatigue in different positions (e.g., how does a player's position contribute to their fatigue?).

General Advice

The challenge is deliberately large and vague. You should feel welcomed to identify a small problem within this much larger problem and even to examine only a subset of the data (e.g., a single game or a single tournament).

Where to Begin

- Watch the video introduction from this year's data providers, the Canadian Sport Institute Pacific.
  - YouTube: ASA Datafest CSI Pacific Introduction Video
- Read the Data Overview which can be found below.
- Read the Data Codebook.
- Consider learning about Rugby 7's.
  - Watch a brief explanation of the rules of Rugby 7's.
    - YouTube: The Rules of Rugby Sevens (Rugby 7's) - EXPLAINED!
  - Watch some highlights, the best "tries," from the 2018 World Cup Sevens.
    - YouTube: Best tries from the women's Rugby World Cup Sevens
    - One of these clips is "in" this year's dataset!
  - Watch a rugby 7's match from this dataset!
    - YouTube: Women's 7s Sydney 2018 Russia vs Canada

Questions?

Provided in the file `datafest-2019-q&a.pdf` are answers to questions about the data from already completed DataFests. Check if you can find an answer to your question there.

Got a question about the data that isn't answered in that file? Send an email to dalpiaz.14@osu.edu with the subject line `[DataFest @ OSU] Data Inquiry` and we'll try to get you an answer!

Data Overview

Common Characteristics

The data were collected during the 2017 - 2018 season. There are five files that give different aspects of the games. The data themselves were collected through a variety of means.

- Player level data are provided by the individual athletes themselves and by IMU / GPS devices worn on their vests during games. GPS data may not be available if players are out of range of the satellites. Players are uniquely identified by the `PlayerID` variable in all data files. Note that players numbered 18 - 21 did not play in any of the games in this dataset, and so they can be removed from the analysis.
- Data are available on each game played during the season. Games are often organized in tournaments, which consist of up to 6 games. Each game consists of two 7-minute halves (except the final game of a tournament, game 6, which consists of two 10-minute halves). Games can have extra time at the referee's discretion, if play is stopped for some reason during the game. There can be up to three games played on a single day. The order and time of the games is provided.

There were a total of 43 games, and they are identified by the `GameID`, which indicates the order in which the games were played throughout the season. (`GameID = 1` is the first game played in the season.)

The data is contained in four different files.

`games.csv`

Tells you when, where, opponent, and high-level outcomes and events in the game ("box scores").

- How collected: Information was pulled from Wikipedia.
- How to use: high-level game information.
- Links: `GameID` links to `gps.csv`. `Date` links to `wellness.csv`, `rpe.csv`

`wellness.csv`

Self-reported health and wellness for each player.

- How Collected: self-reported by each athlete. In principle, reported every morning before 8:30 AM. All values are subjective, but Urine Specific Gravity (USG) is recorded through a sensor. Each athlete may have a different sense of what "typical" means for them, so consider standardizing per athlete.
- How to Use: provides subjective sense of energy levels. USG can provide evidence of dehydration.
- Links: `Date` links to `games.csv`, `rpe.csv`. `PlayerID` links to `rpe.csv`, `gps.csv`.

`rpe.csv`

Rate of **P**erceived **E**xertion. Self-reported workloads for each "session". A session can be a workout (focusing on a particular objective) or a game.

- How Collected: In theory, each player rates herself after each session or game. It is easy, however, for players to neglect this when playing back-to-back games. Note that each day there can be multiple "sessions", and that a "session" can be a recovery period, a game, strength and conditioning, etc. There is no way to associate a particular rating with a particular game on days in which multiple games were played.

- How to Use: Can be used to provide a subjective sense of fatigue. Note that what one player rates "4" for RPE another might rate "7" or any other number, so consider standardizing per player. For many sports analysts, a ratio of acute to chronic training load greater than 1.2 indicates that the athlete is currently in "high" training load and at an increased risk for injury. A ratio less than 0.8 indicates that they are "de-training" or recovering. These are cut-off values based on Australian Football League players.
- Links: `Date` links to `wellness.csv`, `games.csv`. `PlayerID` links to `wellness.csv` and `gps.csv`.

`gps.csv`

Position data for each player during a game.

- How Collected: Data collected from sensors worn by players. Originally, data were collected at 100 Hz (100 times per second), but have been collapsed to 10 Hz. Thus, each second, there are 10 "frames" that provide information on player location and acceleration. Note that we do not know the location of the ball, or the orientation of the playing field. The "z" acceleration is in the up-down direction, "x" is back-front, "y" is side-to-side.
- How to Use: **With caution.** *Note that making plots of location is unlikely to help you understand the role of fatigue unless you first think carefully about aspects of location that might be affected by fatigue.* Some large-scale things to consider: can you infer tackles? Coaches usually encourage players to keep space between them.
- Links: `PlayerID` links to `wellness.csv`, `rpe.csv`. `GameID` links to `games.csv`.

Judging and Awards

Prizes will be awarded in the following categories. At the judges' discretion, honorable mentions may be given in each category. Teams are limited to earning a single distinction (win or honorable mention).

- Best **Insight**: For the team that provides an interesting insight that is supported by the *highest quality analysis of the data.*
- Best **Visualization**: For the team that creates the most compelling and useful visual representation of an aspect of the data or data analysis.
- Best Use of **External** Data: For the team that most effectively uses external information to improve the quality of an insight, analysis or visualization.

In addition to the three main categories, a Judge's Choice award may also be given.

- **Judges' Choice**: Awarded at the judges' discretion to a team whose analysis excels on some dimension not covered by the other award categories, or to the team who is a strong runner-up in one of the three named prize categories.

Advice

Here is some advice that might help you plan your time during DataFest:

- First, make sure you can load the data into whatever software you plan to use. If you run into trouble, look for a mentor, or ask other students. This is meant to be a friendly competition!
- Once you know you can load and access the data, take some time to learn about the data sets. This may take an hour or so, but developing a good understanding of the data is essential to doing a good analysis. You might want to:
  - For each data set, go through the code book variable by variable. While you're doing that, look at typical data values and make some exploratory plots so you have a good understanding of what the data look like.

  - Consider assigning an individual team member responsible for investigate the variables in each of the datasets.
- Now that you have a reasonable understanding of the what the data represent, start thinking about possible questions you might want to answer. You can refer to the suggestions above, but feel free to be creative based on what you see in the data.
  - This process will lead you to start looking at relationships between variables. Think about what summaries of the data and what plots you might want to construct. You might also think about new variables that can be constructed based on the data, or ways of collapsing the data set that might prove useful.
  - Also think about whether you can collect some external data to help you in your analysis.
- By late afternoon / early evening Saturday you should have a good idea about what you want to focus on.
- The rest of the time can be spent refining your analysis (when feasible moving from descriptive statistics to statistical modeling or advanced analysis techniques), creating compelling visualizations, and organizing your arguments for your presentation.

Remember, stay focused and relaxed, ask for help when needed, and HAVE FUN!!

Presentation Submission

Each group must submit a `pdf` of their presentation by **11:59 AM on Sunday**. Please name your file `your_team_name.pdf`. (Or get as close as possible. Some of you used some interesting team names with some characters that might cause problems with filenames...)

- Submit your `.pdf` file at: http://go.osu.edu/datafest-submit

Reminders and additional instructions about submission will be sent via email during the competition.
