# Somersaulting to Gold:
# Optimizing USA's Artistic Gymnastics Team Selection Strategy

Emily Zhao, Elyse McFalls, Young Jun, Evan Kim

2023-11-20

## I. Introduction

Gymnastics has swiftly risen to become one of America's cherished pastimes, with recent polls revealing it to be the most anticipated Olympic event among Americans.[1] This heightened enthusiasm is largely attributed to the historical success the US has enjoyed in the Olympics, particularly in women's artistic gymnastics. Dominant performances by teams like the Final Five in 2016 and athletes like Simone Biles (2011 - current) have brought immense prominence to the sport and its athletes.[2] While the US men's artistic gymnastics team may not be as dominant, they have consistently secured top-five placements in team events at the Olympics and have bolstered decorated athletes like Khoi Young (2018 - current) in recent years.[3]

We expect that millions of Americans are anticipating the gymnastics events at the 2024 Olympic Games in Paris, France.[4] This anticipation is coupled with the expectation of US dominance at the Olympics, mirroring the success we have seen in the past. One may ask then: How can we ensure that the US reigns supreme? The answer lies in assembling formidable teams capable of winning the most medals.

Historically, US Olympic team rosters are shaped by the results of the US Olympic Team Trials. The trials will be hosted in June, 2024, for this upcoming cycle.[5] Here, each invited athlete's performance determines their spot on the team. Athletes are required to complete all apparatuses twice (4 for women, 6 for men), and scores from each event (8 for women, 12 for men) are combined to generate a final score. The selection committee then uses a combination of the competition results and the teams' needs to decide on the final rosters. In the last Olympic cycle, the top two performers for women and the top performer for men at the trials secured automatic spots, while the remaining slots were determined by the selection committee.[6] Given this meticulous process, it will be some time before the US Olympic gymnastics teams are unveiled, and even longer before we can assess how they will perform on the big stage. The selection committee also faces the arduous task of determining the optimal team makeup. This season, the challenge is particularly difficult for the women's team given the abundance of strong candidates.[7]

In this paper, we present a simulation-based approach to determine the best US Olympic men and women's gymnastic teams for the 2024 Paris Olympics. Our method utilizes historical data on each athlete's performances to simulate pseudo-Olympic data to see how well the US will fare under different team makeups. We imagine that our algorithmic approach would be useful supplementary material for the selection committee to consider when finalizing the roster. Moreover, the results of these simulations may also be enjoyed by fans of the sport who are interested in gaining some insight into what to expect from the Olympics given the athletes' past scores.

## II. Methodology

### *Data Collection:*

We were originally provided two cleaned data sets from UCSAS: the 2020 Tokyo Olympics women's results and the 2022-2023 season's men and women's results from various competitions. Given the dynamic nature of gymnastics, where athletes' careers are typically short and marked by rapid changes, the period between two Olympic cycles is crucial. It often sees the retirement of established gymnasts and the rise of new talents, significantly altering the sport's competitive hierarchy. (Atikovic et al., 2017)[8] Therefore, we have elected to only consider the 2022-2023 season's results as relevant data as it more accurately reflects the current competitive landscape and athlete performance trends. In addition, via web scraping and research, we were able to add athletes (both US and non-US) that were not in the provided data set but are worth considering in our simulations.

Any scores that were missing or recorded as "0" were dropped from the input data for the simulation as we assumed the athlete had chosen not to compete in the event. We also extensively cleaned the data for inconsistencies, mis-typed, and changed names for each athlete, while also checking for athletes who switched country allegiances.

***Motivation for Kernel Density Estimation (KDE):***

Our refined and aggregated data set reveals the distribution of scores for each athlete-apparatus combination over the most recent two years. This observation suggests that each combination possesses a unique, true underlying distribution of scores, justifying the need for density estimation. Our method of choice is Kernel Density Estimation (KDE), as opposed to piece-wise histogram density estimation, given the near-continuous nature of gymnastics scores. It offers a significant advantage by centering a smooth kernel function at each data point and aggregating these to form an overall density estimate.

While there are over twenty packages in R that assist with density estimation, varying in both theoretical backdrop and computational performance, we employ the `KernSmooth::bkde()` function, renowned for its speed, accuracy, and robust maintenance. Developed by Matt Wand and Brian Ripley in 2010, `KernSmooth::bkde()` implements the linear binning technique, which distributes data mass between grid points for enhanced accuracy by utilizing "weighted bin counts instead of individual data points". (Deng et al., 2011)[9]

***Developing the Simulation:***

For the purpose of our analysis, we have fine-tuned the `KernSmooth::bkde()` function to our data set's specific characteristics. We selected the Gaussian kernel for its efficacy in creating a smooth, continuous approximation of the underlying score distributions, while deliberately forgoing the scale normalization (by setting `canonical=FALSE`) which is typically associated with canonical kernels to maintain the natural scale of gymnastics scores.[1]0 The `bandwidth` parameter is set as 0.25 all throughout, which yields the optimal empirical balance between over-smoothing and under-smoothing for our aggregated data set. (See Appendix, Figure 1)

Furthermore, for the `range.x` parameter, we have employed tailored lower and upper limits as to the range of scores for which we should estimate the density. Initially, the lower limit is fixed at 0, a logical baseline for gymnastics scores, while the upper limit corresponds to the highest recorded score in our dataset, 15.7. Subsequently, we have introduced a `nugget` metric, adding extra leeway, particularly for athlete-apparatus score data with fewer than four observations, thereby providing a buffer that accommodates the potential for higher variability in sparser data. This methodology is grounded in the principle that past performance can be a predictor of future results, but with a recognition of the stochastic nature of athletic performance. The nugget is calibrated using the standard deviation of all scores for a given apparatus, rewarding consistent high-level performance over time rather than exceptional one-off performances.

Incorporating our fine-tuned `KernSmooth` function with our own simulation algorithm, we conducted a total of 1,000 iterations for the following: 4 women's individual apparatuses and 6 men's individual apparatuses. By conducting a total of 1,000 iterations of simulated Olympic events (separately for both men and women), we generated robust data sets reflecting the potential range of scores for each athlete-apparatus combination, assuming all athletes compete in the events they have historic data for. The aggregation of these results facilitated a comprehensive analysis of the top performers for each apparatus and potential team compositions for both the US and other team-qualifying countries.

***Weighted Individual Medal Contributions:***

In order to determine the medal potential of each US gymnast, we calculated their “medal score” for each simulated event. With gold assigned as a value of 3, silver as 2, bronze as 1, and anything lower than third place as 0, we assigned each competitor of each simulated event their “medal score”. When two gymnasts happen to earn the same score within one of the top three spots, they are both rewarded the same medal score. This medal score is naturally relative to the performance of other gymnasts and is a better representation of a gymnast’s potential on the team than their score. Then, across simulations of each event, we found each competitor’s average “medal score”. By conducting this weighted medal calculation for each women’s and men’s event, we can determine the most skilled gymnasts by event and also determine how frequently they appear to medal.

This calculation generalizes from the exact place that the gymnast is predicted to achieve to a range between 0 and 3. For instance, if a gymnast only scores bronze once throughout the 1,000 simulations of an event, their weighted medal score would be 1. Meanwhile, a gymnast who consistently earns silver or gold throughout the 1,000 simulations will obtain a weighted medal score between 2 and 3. This process allows us to identify which individual events Team USA has the highest chance of winning and whether there are particular events that may benefit from a “specialist”, a gymnast who may only excel at one particular event.

We can use this process of calculations for all four apparatus for women, all six apparatus for men, and the individual all-around event for both women and men. The individual all-around event weighted score calculation is derived from the same simulation data as the individual event. The individual women’s and men’s all-around scores are calculated by taking the sum of the athlete’s performance in all 4 and 6 respective events for each simulation. Then, the same ranking and medal scoring process for each simulation is applied to the summed all-around scores.

***Weighted Individual Medal Contributions:***

From the apparatus individual event simulations, we slice for the top three athletes in each apparatus per country to generate a group of top-performing women and men. From this group, we create all possible combinations of 5 athletes from the groups of top performers per country. This left us with all of the possible teams for all of the qualified countries given our criteria. Then, following the team all-around rules and using the top three scores for each event, we generated all of the scores for each team.[1]1 The team with the highest score for a given country was selected for each simulation and ranked against the other countries’ best teams to determine medal placements for the simulation.

For men’s, the other qualifying teams include Canada, China, Germany, Great Britain, Italy, Japan, Spain, Switzerland, Netherlands, Turkey, and Ukraine. For women’s, the other qualifying teams include Australia, Brazil, Canada, China, France, Great Britain, Italy, Japan, Netherlands , Romania, and South Korea.[1]2 All these teams had at least five athletes qualify under our initial threshold criteria except for the men’s Netherlands team which only had four athletes qualify under our criteria. After further research on the historic performance of the Netherlands men’s team, we believed that they are not in contention for a team all-around medal and therefore removed their team from the team simulations.[1]3

***Assumptions:***

We applied a series of assumptions to reduce noise in the data, computational demand, and simplify the selection process. Firstly, we decided to establish a performance cutoff for the competitors inputted into the simulation function. We removed any athlete who has not placed in the top 5 for any competition in the 2022-2023 seasons, as we assume that their either lack of data from the recent seasons or their relatively sub-par performance would essentially eliminate their possibility for qualifying for the Olympics. This criteria is supported by the 2020 Team USA artistic gymnastics team selections as all members of the selected teams had placed in the top 5 of a competition from 2019 to 2020. This criteria allows for flexibility as even if a gymnast is a “specialist”, their top 5 strong performance in a singular event will be sufficient for including them in the data set.

In the selection of our team, we also assume that these gymnasts will be performing at their optimal condition. In our selection process, we have also removed some athletes which have either announced their retirement or have suffered a recent significant injury. One notable gymnast who was removed from the data set due to this criteria is Suni Lee. At her peak performance, gymnastics fans believe she's a strong contender on both uneven bars and balance beam.[1]4 However, in April 2023, she announced that she was withdrawing from collegiate gymnastics and had been diagnosed with an incurable kidney disease. The uncertainties surrounding her disease and its progression over the next year would potentially make her chance of performing well volatile, so we could not confidently place her in our recommendation given her medical condition unknowns. For the rest of the athletes, we assume that gymnasts will be able to participate in multiple events throughout the Olympics competition and will also not get hurt or injured from now until the Olympics.

In the simulation of results, we grouped all historic vault scores, either from team or the first or second round of individual vaults, all into a common vault category, unifying labels such as "VT1" and "VT2" into simply "VT". We also applied the assumption that a gymnast's performance in each round - qualifying, final, or all-around final - would all follow a similar kernel density distribution and therefore can be aggregated into one cumulative apparatus category.

## III. Results

## IV. Discussion

***Limitations:***

***Extensions:***

## V. Appendix

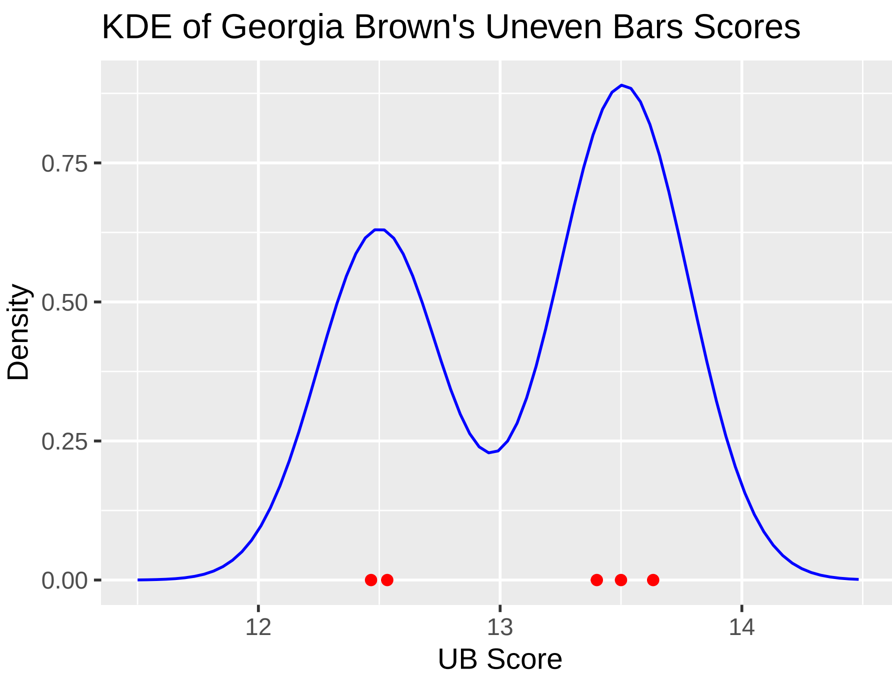


Figure 1: *Above is an example of the KDE density estimation for a specific athlete-apparatus combination based on scores from the 2022-2023 seasons, using our fine-tuned `KernSmooth::bkde()` function. The red dots represent the scores from the data set, and the blue line estimates the underlying distribution, using the Gaussian family as a parameter.*

## VI. Sources Cited


1. "Poll: Gymnastics Is America's Most Highly Anticipated Olympic Sport". SGB Media. 2021. https://sgbonline.com/poll-gymnastics-is-americans-most-highly-anticipated-olympic-sport/
2. "How one adjustment helped USA Gymnastics become a power and led to Rio gold". Sports Illustrated. 2016. https://www.si.com/olympics/2016/08/10/usa-womens-gymnastics-system-rio-olympics-martha-karolyi.
3. United States men's national artistic gymnastics team. Wikipedia. https://en.wikipedia.org/wiki/United_States_men%27s_national_artistic_gymnastics_team
4. Paris 2024 - Artistic Gymnastics. https://www.paris2024.org/en/sport/artistic-gymnastics/
5. "Minneapolis, Minnesota to host 2024 US Olympic Team Trials – Gymnastics, become Gymnastics City USA 2024". USA Gymnastics. 2023. https://usagym.org/minneapolis-minnesota-to-host-2024-u-s-olympic-team-trials-gymnastics-become-gymnastics-city-usa-2024/
6. "What to know about the US Olympic gymnastics trials". The Seattle Times. 2021. https://www.seattletimes.com/sports/olympics/what-to-know-about-the-u-s-olympic-gymnastics-trials/
7. "Picking a strong US gymnastics team is easy. Whom to leave out is hard". The Washington Post. 2023. https://www.washingtonpost.com/sports/olympics/2023/08/29/us-gymnastics-team-world-championships/
8. Atikovic, A., Kalinski, S. D., Cuk, I. Age trends in artistic gymnastic across world championships and the olympic games from 2003 to 2016. Science of Gymnastics Journal. 2017. https://www.researchgate.net/publication/320335292_Age_trends_in_artistic_gymnastic_across_world_championships_and_the_olympic_games_from_2003_to_2016
9. Deng, H., Wickham, H. Density estimation in R. 2011. https://vita.had.co.nz/papers/density-estimation.pdf
10. Wand, M., Moler, Cleve., Ripley, B. "Package KernSmooth". CRAN. 2023. https://cran.r-project.org/web/packages/KernSmooth/KernSmooth.pdf
11. "Gymnastics 101: Competition format". NBC Olympics. 2021. https://www.nbcolympics.com/news/gymnastics-101-competition-format
12. "Who Qualified to Paris 2024?". The Gymternet. 2023. https://thegymter.net/2023/10/09/who-qualified-to-paris-2024/
13. "Worlds Team Preview #8 - The Netherlands". Reddit, r/Gymnastics. 2023. https://www.reddit.com/r/Gymnastics/comments/16nqb9r/worlds_team_preview_8_the_netherlands/
14. "Olympic gymnastics champion Suni Lee began pointing to 2024 long ago. It could be a difficult road". AP News. 2023. https://apnews.com/article/sunisa-lee-paris-olympics-gymnastics-95c3857ec9338170792baca56ffe6c07