Greatness in Sports:
Understanding and Visualizing its Patterns

Benjamin Leinwand and Matthew Murray

Note: Limitations/concerns are noted in red.

We are aware and cognizant of the fact that if a reader were to conduct a similar analysis, the results and data may be very different. However, our goal in this project is not to conduct an analysis that is extremely robust/resistant to sensitivty tests, but rather, one that is informative, insightful, and interesting. If 25 different researchers were to conduct this analysis, each dataset and approach may be very different, and we are okay with that fact.

1 Introduction

1.1 Background

A common phrase that is emphasized among sports fans and expert calls to not take greatness for granted. While the sports themselves are (hopefully) lasting, the athletes come and go, with their careers seeming like a brief blip in time when it is all said and done. Baseball fans listen to stories of the Great Bambino from their elders, wondering if they will ever see a player or a talent burst on to the scene again. Meanwhile, basketball fans seem to engage in a persistent dispute between Michael Jordan and Lebron James.

Nonetheless, with the seemingly endless abundance of talent on display today, coupled with the advances being made in athlete performance and recovery that have athletes performing at the highest levels ever, one has to call to question: can we take greatness for granted? Are there more greatest of all time (GOAT) candidates than one may intuitively believe there to be? Are there patterns in which greatness occurs in one sport, and a lack thereof in another? Do certainly qualities (e.g: contact vs. non-contact) tend to attract more GOAT candidates to a sport than to another. In this paper, we seek to provide the answers to these questions.

1.2 Data Collection

1.2.1 Sports

Due to the unique nature of this project, we had to impute our own datasets. We decided to compile lists of mens GOAT candidates from the following sports:

1. Soccer
2. Cricket
3. Tennis
4. Basketball
5. Baseball
6. Football
7. Ice Hockey
8. Golf

The inclusion of football, baseball, basketball, and ice hockey can be attributed to the fact that they are the 4 of the 5 most popular sports in the United States (Gallup 2017). Golf and tennis were included because professionals in both sports compete in four "major" tournaments each year. Soccer and cricket were included due to their immense popularity and following worldwide (Media 2017).

1.2.2 GOAT Sources and Rationale

For each sport, we imputed two lists that rank the best players of all time. One list would be a subjective ranking of the players from a reputable, respected source. While the reputability of a source is inherently subjective, we tried as much as possible to include sources that are mainstream and whose respectability would be difficult to refute. The sources of the subjective ranking lists for each sport are as follows:

- Soccer - *Sports Illustrated*
- Cricket - *British Broadcasting Company (BBC)*
- Tennis - *Stadium Talk*
- Basketball - *ESPN*
- Baseball - *ESPN*
- Football - *ClutchPoints*
- Ice Hockey - *Bleacher Report*
- Golf - None

The other list would be based on an objective metric, such as the number of MVP's won throughout one's career or a comprehensive statistic that aptly captures a players value. We tried to use metrics that are both fair and comprehensive in that they thoroughly and aptly capture a player's greatness. The criteria/measures used for each sport - as well as the inevitable limitation(s) of each criteria - are noted below:

- Soccer - players were ranked by the number of Ballon d'Ors they won. More specifically, the list includes the number of Ballon d'Ors won by each player *since 2007*. The Ballon d'Or was chosen because it is widely recognized as the most prestigious award in soccer. Nonetheless, this award does have a few limitations. From 1956 (the first year the award was given out) to 1995, the Ballon d'Or was only given to the best soccer player of European origin. In 1995, the award was expanded to consider soccer players from any European club. It was not until 2007 that the award was expanded to consider all soccer players worldwide (Molinaro 2011).
- Cricket - Players were ranked based on their batting average. We arbitrarily chose batting average because it is one of the most commonly cited offensive statistics for cricket. (For instance, Donald Bradman is often noted as the greatest cricketer of all *because* of his ridiculously high batting average.) It is equal to the number of runs a player has scored divided by the number of time he has gotten out; in other words, it is the average number of runs a player scores at bat before getting out.
- Tennis - Players were ranked based on the number of majors they won. The four majors, also known as Grand Slam events, in tennis include the Australian Open, the French Open, Wimbledon, and the US Open, each of which is played annually. These tournaments became open to professionals in 1968, thus marking the advent of the Open Era in tennis as well as a significant shift and maturation of the sport (Crim, n.d.). Today, tennis player accomplishments are often discussed in the context of Open Era records, with one of the most prominent ones being Grand Slam or major titles.
- Basketball - Players were ranked based on their Value over Replacement Player (VORP). This metric measures "the number of points that a player contributed per 100 team possessions above a replacement level player" (Bannon 2022).
- Baseball - Players were ranked based on their Most Valuable Player (MVP) shares. Put simply, an award share is equal to the number of points that a player received for an award divided by the total number of first place votes. For example, assuming there are 10 votes, if a player wins 9 first place votes, each of which are worth 3 points, and 1 second place vote, which is worth 2 points, the player wins up with an award share of $29/30$ (≈ 0.967). The benefit of using this metric is that it rewards and considers players who did not necessarily win the MVP, but were close.
- Football - Players were ranked based on the number of MVP Awards they won. More specifically, the list includes players who won more than 1 NFL MVP.
- Ice Hockey - Players were ranked based on their points. In hockey, a player earns 1 point every time he either scores or assists a goal.

- Golf - Players were ranked based on the number of majors (Masters, PGA Championship, U.S. Open, and The Open) they won. More specifically, we included players who have won at least 5 majors. Additionally, we ranked players based on PGA tournament victories.

1.2.3 Time Span

Each of our graphs includes data/players from 1945 to the present (although we may later change this to 2010; after 2010, we are likely underestimating the number of GOATs just because most of the great players have not been playing long enough to truly be in the GOAT conversation; a perfect example of such a player would be Patrick Mahomes II, who plays for the Kansas City Chiefs). The choice of 1945 can be attributed to the fact that at this time, the MLB (1876) and NFL (1920) were already founded, while the Basketball Association of America (which eventually became the NBA) was founded shortly thereafter in 1946. Therefore, one can argue that the three most popular sports in North America (in addition to soccer) were already up and running by that time period. Additionally, 1945 marked the end of World War II and the advent of a period of economic flourishing in the Western world.

1.2.4 Career Span

Another challenge that we encountered during the data imputation process was properly, consistently, and logically defining defining the beginning and end of a player's professional career. In doing so, we hoped to capture the **entirety** of a player's career, including the advent, peak, and twilight of one's playing years. The specific details of how we defined the beginning and end of a player's career is below:

- Soccer - We imputed a player's active years as the first and last calendar years that one played professionally, either internationally *or* for a professional club), and all years in between.
- Cricket - We imputed a player's active years in 1 of 2 ways. If the player was from our arbitrary list, the active years equate to the first and last calendar years that one played professionally - either internationally *or* domestically - and all years in between. If the player was from our list based on batting average, the active years equate to the first and last year he played in a test match, and all years in between. In the case that a player is on both lists, we used the former method.
- Tennis - We imputed a player's active years as the first and last calendar years that one played professionally
- Basketball - We imputed a player's active years as all years that he played in the NBA.
- Baseball - We imputed a player's active years as all years that he played in the MLB.
- Football - We imputed a player's active years as all years that he played in the NFL.
- Ice Hockey - We imputed a player's active years as all years that he played in the NHL.
- Golf - We imputed a player's active years as the first and last calendar years that he won a major, and all years in between. For example, since Tiger Woods won his first major in 1997 and his last major in 2019, his active years are listed as 1997-2019, even though he is still playing in professional golf tournaments (in fact I know he most recently played in the Genesis Open). The reason why we used this logic is that professional golfers can play for a longer (compared to atheletes in other sports) period of time due to the presence of the PGA Tour Champions (formerly known as the Senior PGA Tour) as well as the nature of the sport itself (more specifically, the fact that players can sit out certain tournaments and the less violent nature of the sport). For example, Jack Nicklaus, whom many consider to be the greatest golfer of all time, played at his last U.S. Open in 2000, but did not play his last Masters, PGA Championship, and Open Championship until 2005 (add source here).

Another limitation to take into account is the fact that we imputed data year-by-year for athletes in some sports, but not in others. A potential result of this limitation is that said years may have an inflated number of active GOAT candidates.

Topics of Interest

Contact vs. Non-Contact Sports

One distinction that we chose to investigate is contact sports vs. non-contact sports. An article published by *The American Academy of Pediatrics* titled "Medical Conditions Affecting Sports Participation" divides sports into three main categories: (1) contact, (2) limited-contact, and (3) non contact. Furthermore, contact sports are sub-divided into (1) collision and (2) contact sports. It further states:

"In collision sports (eg, boxing, ice hockey, football, lacrosse, and rodeo), athletes purposely hit or collide with each other or with inanimate objects (including the ground) with great force. In contact sports (eg, basketball and soccer), athletes routinely make contact with each other or with inanimate objects but usually with less force than in collision sports. In limited-contact sports (eg, softball and squash), contact with other athletes or with inanimate objects is infrequent or inadvertent. However, some limited-contact sports (eg, skateboarding) can be as dangerous as collision or contact sports. Even in noncontact sports (eg, power lifting), in which contact is rare and unexpected, serious injuries can occur."

Using the criteria noted above, we decided to classify **basketball, football, hockey, and soccer** as **contact** sports; conversely, **baseball, cricket, golf, and tennis** were deemed **limited contact or non contact**. In doing so, we were looking to investigate whether the type of sport that a player participates in affects a GOAT candidate's career length.

—(Include Figure 11 here)—

Based on **Figure 11*, no trends in particular seem to stick out. Although football and basketball (both of which are contact sports) both seem to have shorter average career lengths, one cannot say this for all contact sports compared to non-contact sports. However, a potential limitation/inconsistency for golf (a non-contact sport) is that the average career length may appear to be shorter than it actually is because we used the first and last years a player won a major for a career span.

Maximal Value for Increasing Distribution

Let's examine this from a probabilistic perspective. Consider a standard uniform distribution, that is randomly picking any real number between 0 and 1 with equal probability, including any fraction or decimal in that range. Imagine we keep drawing random numbers from this standard uniform distribution. The first value we pick must be the largest value we have observed as of yet, just like the first competitor to complete a time trial will always be the leader in the competition, even if only temporarily. However, the second value we draw has to be larger than the first value, and since they're both drawn randomly, it has probability ½ of being the largest value. The 3rd value would have to be larger than the first and second values, which has probability 1/3. This logic continues, meaning the probability that the nth draw is larger than all previous draws is 1/n. Under that model, we expect that we will see a lot of "bests" early on, but fewer and fewer as time goes on. This makes intuitive sense. When a sport is new, an athlete is competing with only a few predecessors. Later on, when competing to be the best ever, an athlete may be competing with the same number of contemporaries, but also competing against many more "ghosts" who could have finished their careers decades ago.

This simple model assumes that the distribution remains the same, in other words that there is no growth over time. Let's look at the other extreme that holds in a completely different domain. Moore's Law states "the number of transistors on integrated circuits doubles about every two years." In that case, as long as Moore's Law holds, we systematically expect the latest generation of computers to be much faster than all previous generations. It's worth noting that by invoking doubling, Moore's Law describes exponential growth, but even linear growth could suffice to expect new computers to be faster than old computers. Consider a variant of Moore's Law that would instead say "the number of transistors on integrated circuits increases by 1000 every two years."

This linear variant of Moore's law might describe the progress of professional athletes. Over time, sports has become a bigger business, and training methods have improved. J.J. Redick caused a bit of a stir when he said Bob Cousy played against "plumbers and firemen." As a stylized model, let's imagine the first year, players are drawn from a N(0, 1). In any subsequent year i, the players are drawn from a N(c(i-1), 1). This means that the expected average ability increases by c units each year. If c is very large, we'd expect each year to be much better than every previous year. This is still a very simple model, that assumes a constant linear increase in skill, when in fact, skill growth may accelerate or decelerate. Maybe rule changes, or new equipment, or revolutionary strategies represent a larger one-time spike. It seems plausible that new swimsuits in 2008-2009 represented a one-time drastic boost, but that advantage was soon removed when the suits were banned, leaving the original moderate improvement trend without the spike from equipment. Furthermore, improving competition may obscure greatness, as an outlier in an earlier era may be merely excellent in a later one.

Let's examine this simple model. Imagine that each year for 100 years, n new athletes turn professional. Each one of those n new professionals has the same probability of exceeding all values observed in prior years, given by:

$$P_{i1} = \int_{-\infty}^{\infty} \prod_{j=1}^{i-1} \prod_{k=1}^{n} F_{x_{jk}}(a) f_{x_{i1}}(a) da$$

Where F_(X_jk) represents the CDF of a value in year j, and f_(X_i1)represents the pdf of a value in year i. The probability that no value in year i is greater than all values observed in previous years is therefore (1- Pi1)n , so the probability the greatest year appears in year i is 1- (1- Pi1)n.

Chart 1 shows the probability that under different values of c and n, an athlete debuting each year will have the highest value observed so far.

n = 10, 100, 1000 c = .01, .05, .1, 1

Competition complicates things

Revolutions in Sports

Background

Various changes in rule, equipment, and technique have enabled athletes to gain competitive advantages over their predecessors in a plethora of ways. We now elucidate on some of the most prominent ones in the sports that we have covered.

1. Fosbury Flop

The Fosbury flop is the most predominant technique used in the track and field event high jump. It was introduced by US Olympian and gold medalist Dick Fosbury, who first used the technique at the 1968 Summer Olympics in Mexico City. The technique was revolutionary in the sense that it consists of jumping backwards off the "wrong" foot. While Fosbury was the first athlete to use this technique in 1968, it was quickly adopted, as 28 of the 40 high jumpers used it in the 1972 Munich Olympics. Furthermore, the technique was quickly deemed to be the most effective and efficient way to perform a high jump. Today, the Fosbury flop is the only technique used by Olympic level high jumpers.

The Fosbury flop is a prime example of how advancements in technique and strategy can give athletes advantages over their predecessors. The advent and introduction of the technique has been linked to the rapid world record progression of the high jump in the 1970's and 80's. More information about it can be found here(Minshull 2018).

2. Tech Suits

Tech suits burst onto the professional swim scene shortly thereafter the 2004 Summer Olympics in Athens when Speedo asked NASA to help it design swim suits will less drag. In the subsequent yeats, Speedo released the LZR Racer, which furthermore led to the release of various other aerodynamic tech suits - or "susper-suits" - from prominent swimwear companies. During this period - known as the "shiny suit era" - swimming world records were broken at an unprecedented pace. More specifically, *74* world records were broken by swimmers wearing the LZR Racer suit. Consequently, World Aquatics, formerly known as the Federation Internationale de Natation (FINA), controversially banned tech suits in 2009, which has led a noticeable drop-off in the pace at which world records have been broken (McCluskey 2019).

Tech suits are a prime example of how advancements in equipment can give athletes advantages over their predecessors. More information about them and the impact they had on world records can be found here(Filocca 2018).

References

Bannon, Chuck. 2022. "NBA MVP - Digging into the Numbers." https://www.qlik.com/blog/nba-mvp-digging-into-the-numbers; Qlik.

Crim, Jon. n.d. "Open Era in Tennis, 1968." https://tenniscompanion.org/open-era-in-tennis/#open-era-records; Tennis Companion.

Filocca, Giulia. 2018. "Can World Records Supersede the Super-Suit Era?" https://www.swimmingworldmagazine.com/news/can-world-records-supersede-the-super-suit-era/; Swimming World.

Gallup. 2017. "Sports." https://news.gallup.com/poll/4735/sports.aspx; Gallup.

McCluskey, Lianne. 2019. "One Decade Later, Do We Miss the Full-Body Tech Suit?" https://www.swimmingworldmagazine.com/news/one-decade-later-do-we-miss-the-full-body-competition-suit/; Swimming World.

Media, News Members, News Freeview. 2017. "The World's Most Watched Sports." https://sportforbusiness.com/the-worlds-most-watched-sports/; Sport for Business.

Minshull, Phil. 2018. "50 Years Since the Day Dick Fosbury Revolutionised the High Jump." https://worldathletics.org/news/feature/dick-fosbury-flop; World Athletics.

Molinaro, John F. 2011. "History of the Ballon d'or." https://www.cbc.ca/sports/soccer/history-of-the-ballon-d-or-1.899151; CBC.