12-2015

Analytics in the ATP

Christopher Pillitere
Clemson University, cpillit@g.clemson.edu

ANALYTICS IN THE ATP

A Thesis
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Master of Arts
Economics

---

by
Christopher Pillitere
December 2015

Accepted by:
Dr. Raymond Sauer, Committee Chair
Dr. Curtis Simon
Dr. F. Andrew Hanssen

ABSTRACT

This paper applies data analysis used in the major team sports to the Association of Tennis Professional (ATP). A linear regression model is used to determine the aspects of the professional tennis most relevant to player success. Special focus is given to the effect of a player's overreliance on his serving performance. The available data (1991-2014) is divided into two periods in an effort to observe any differences between the era dominated by Pete Sampras and the era dominated by Roger Federer. The results from the former period indicate the same effect on player success of first and second serving and returning and an insignificant effect of reliance on success. Results from the latter period appear to show a greater importance for first serves and returns and a significant negative effect for reliance on performance.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

THE MOTIVATION

Baseball writer and historian Bill James is widely considered to be the pioneer in the use of empirical analysis of professional baseball, commonly referred to as sabermetrics. An avid lifelong fan, James began writing articles in the 1970s. This would lead to the self-publishing of an annual titled *The Bill James Baseball Abstract*. The book contains in-depth statistical analysis from James's study of box scores. A key contribution by James mentioned in the book is the concept of runs created by hitters in baseball. As James notes: "A hitter's job is to create runs for his team. A hitter is not at the plate attempting to compile a high batting average, or a high slugging average, or high total average, but rather to create runs for his team."[1] Billy Beane, general manager of the Oakland Athletics, piggybacked off this notion of needing offensive players to create runs to benefit his organization. The success of the Oakland Athletics in the early 2000s with a relatively low payroll has led to a revolution in professional sports, spurred on by the release of Michael Lewis's now-famous *Moneyball*.[2] At the heart of Oakland's good performances was the ability of Billy Beane to exploit inefficiencies in the labor market for professional baseball talent. As a result, Beane was able to acquire more on-the-field productivity for less money than his competitors. In essence, the Athletics were getting more bang for their buck. This allowed them to compete with big-spending clubs from larger markets, such as the New York Yankees and Boston Red Sox. In 2001 and 2002 for example, Oakland had near the top records in Major League Baseball while having payrolls around half the league average. The A's were consistently operating on the efficiency frontier, obtaining far higher winning percentages than others organizations in Major League Baseball with similar payrolls.[3]

A key conclusions of Oakland's analysis was the need to place emphasis on on-base percentage. Relative to other statistics, on-base percentage was being undervalued by MLB front offices, thus allowing room for exploitation by Billy Beane. The intuition behind Oakland's findings is in fact remarkably simple. Getting on base leads to more opportunities to score runs, thus increasing the expected number of runs in a given inning, game, or season. For this reason, on-base percentage is to be valued over some of the other metrics of offensive performance in baseball (like batting average, for example). By 2003, the rest of Major League Baseball had caught on.

Analytics has been explored extensively in basketball as well. In his book, *Basketball on Paper*, Dean Oliver explores the best offenses and defenses in basketball. He begins by giving the reader a point of reference, noting that "average is boring, but you need to know what average is in order to define greatness."[4] Oliver is careful to mention the fact that average is a term relative to a team's era. For example, an average offense in the NBA in 1974 would have been considered a bad offense in 1984. According to Oliver, we should "evaluate the teams of history against the averages of their times." His research and experience led him to four critical aspects of the game: shooting percentage from the field, getting offensive rebounds, committing turnovers, and going to the foul line a lot (and making the subsequent free throws). Teams who excel in these categories are the ones who achieve greatness. Conversely, a team who struggles in one of these categories better be able to make it up in the other categories, otherwise the team is likely to be bad.

Today, analytics of this nature is commonplace in all of the major professional team sports. As a native Houstonian, I have seen in recent years what appears to be the positive effects of analytics for the Rockets of the National Basketball Association. Rockets' general manager Daryl Morey is widely regarded as the NBA's leading stats guy. I am hopeful Morey's work will lead the Rockets to a championship title, something that has eluded Billy Beane and the Oakland A's.

Despite its widespread use in professional team sports, Moneyball-style analysis has gained less traction in individual sports. The reasons for the lack of an analytics movement in individual sports have become clearer recently.[5] For one, being an individual sport makes it fundamentally different than baseball and basketball. In the early 2000s, Billy Beane successfully exploited labor market inefficiencies in order to stretch their limited payroll as far as possible. Beane had to assemble 25-man and 40-man rosters, not to mention the development of a strong farm system. This meant acquiring dozens of professional baseball players during the time frame for which the Athletics were ahead of the curve.

While a vibrant labor market exists for professional team sports, essentially no labor market exists in individual sports like golf and tennis. A key feature of the analytics movement is the ability to isolate an individual's contribution to the team's success. In tennis, there is nothing to isolate. We already know exactly what a player has contributed to his own on-court success. Additionally, a struggling tennis player cannot trade for or sign as a free agent a more competent one to play in his place. He is stuck with himself, with his only option being to improve his own skills. This leads me to the next reason why analytics may not have as great a demand in professional tennis. In professional North American sports, a large degree of parity exists. Since the turn of the millennium, fifteen seasons have produced nine different World Series champions in the 30-team Major League Baseball. Furthermore, the highest team winning percentage during this time period was .716,[6] when the 2001 Seattle Mariners won an unusually high 116 regular season games. The competitive balance in professional baseball allows for even small exploitation of market inefficiencies to result in noticeable differences in team performances. In stark contrast is a total lack of parity is professional men's tennis. Over the past decade, nearly 90 percent of the ATP's four major tournaments have been won by one of three players: Roger Federer, Rafael Nadal, and Novak Djokovic. The total dominance by these top players makes it unlikely that any insights gained from comprehensive data analysis will allow players like John Isner or Gilles Simon to be regular contenders for major titles. In some respects, the best current use of analytics in professional tennis comes in helping players determine which tournaments to enter. For players at the very top, the decision on which tournaments to play is fairly straightforward. However, players slightly further down in the world rankings often must decide between multiple events each week with varying amounts of rankings points and prize money available. The tradeoffs players face in these situations is simple enough. Bigger tournaments come with more prize money and more rankings points, yet also have a tougher field of competitors, meaning a player is more likely to be eliminated from the tournament in the earlier rounds. Conversely, players face less stiff competition in the smaller tournaments, but the rankings points and prize money earnings potential is much lower too. The analytics suggest that players are typically better off when they enter into the biggest tournament available to them. They realize higher expected payoffs in both rankings points and prize money.

Despite its limitations, I still believe that analytics have the potential to have a meaningful impact in professional tennis. The knowledge a player can gain from analyzing the data has the ability to help enough to elevate him to a higher level. After all, the prize money available in the larger tournaments is quite significant, and if using analytics allows a player to win just one or two more big matches in a given year, the effect this would have on prize money and year-end ranking is enormous.

I have been an avid follower of men's professional tennis since the early 2000s. My initial interest in the game coincided with the rise of Andy Roddick into the upper echelons of the professional ranks. The sense of American pride I have has led me to great interest in international competitions such as the FIFA World Cup and the Olympics. For me, tennis has been no different. I thoroughly enjoyed watching Andy Roddick throughout his career play against the best from other nations. Of course, over the years I have gained a tremendous amount respect for the talents of the Roger Federer and Rafael Nadal. In my many years of observing the world's top professional tennis players, the stark contrasts in playing style has been very clear. This made me wonder, are all styles of play created equal? Which are most effective? Or is raw tennis talent all that matters? These questions led me to reflect on a common refrain uttered by tennis commentators on television. On a regular basis over the years, I have heard television analysts say that the serve is the most important shot in the game of tennis. Additionally, according to tennis coach and author Frank Giampaolo, the "most glaring example of the game's evolution is found in groundstrokes. It can be argued that the serve is still the most important shot and the greatest potential weapon, but the serve has not seen the metamorphosis that forehands and backhands have undergone over the past generation."[7] I would like to analyze the data to see if this really is the case. If not, what aspects of the game do have the greatest impact on player performance? Is it possible that a certain skill is undervalued at the professional level? If in fact it is possible for a player to determine which aspects of the game (serve, groundstrokes, return, etc.) have the largest impact on winning tennis matches and which skills are undervalued, he would find himself in a situation similar to that of the Oakland Athletics in the early 2000s. He would be equipped with more information than his opponents, thus able to use his practice time most efficiently. That is to say, he could budget a greater amount of his time practicing to areas of the game of tennis that will have the largest impact on winning tennis matches.

Tennis has become a very lucrative sport. Roger Federer, a player many consider to be the greatest of all-time, has earned $88.6 million in prize money in his career as of the end of the 2014 season. The difference in prize money between winning a tournament and finishing in second place can be huge. To put things into perspective, in the 2014 US Open, one of the year's four major tournaments, the winner took home $3 million while second place received $1.45 million. Semifinalists received $730,000 and quarterfinalists received $370,250. One can easily see the difference a win or a loss in a single match can make. If a player can gain a slight edge using statistical analysis, he could potentially walk away with much more money in his bank account at the end of the day.

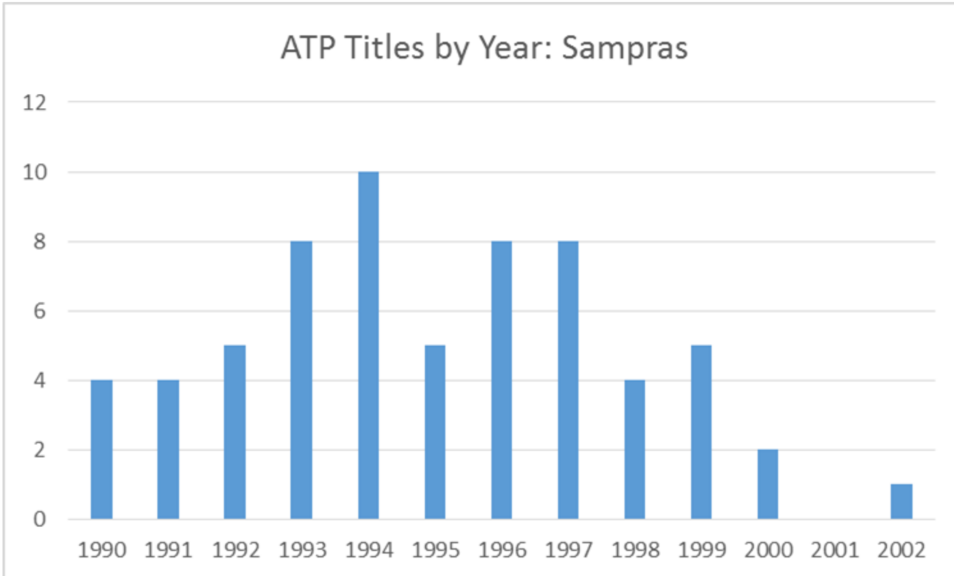
CHAPTER TWO

THE DATA

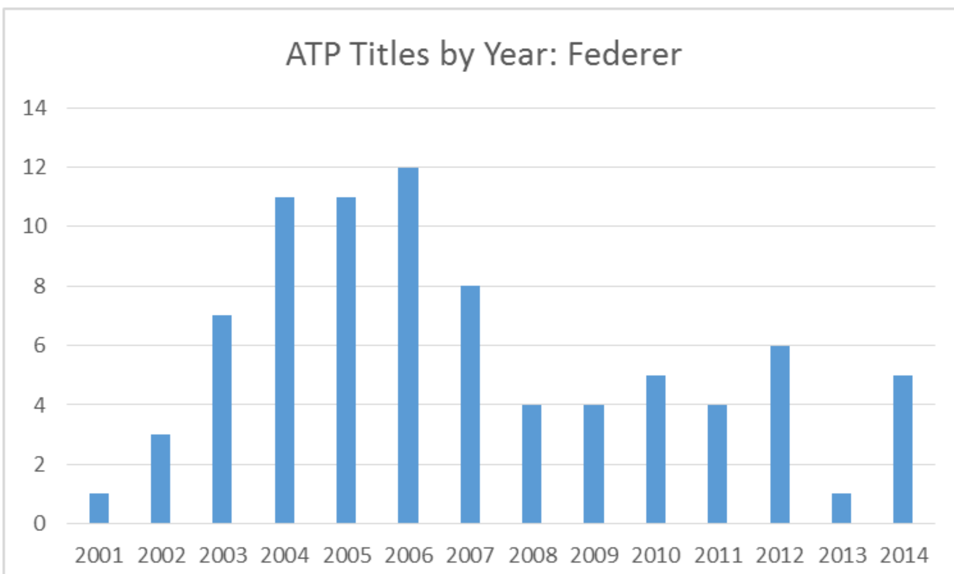
Of all the professional sports in America, it comes as no surprise to me that baseball was the first to experience a revolution in statistical analysis. Apart from being the nation's most popular sport for decades, baseball statistics have been tracked with incredible accuracy for quite a long time. Because of the large amounts of readily available baseball data, the sport has lent itself to rigorous analysis by statisticians and economists alike. Furthermore, the nature of the game of baseball is such that it is easier to differentiate the effects of individual players on a given team's performance. In American football for example, it is difficult to quantify the positive effects of good blocking by offensive lineman or effective clogging of running lanes by a good defensive nose tackle like Vince Wilfork. Likewise, calculating the positive effects on team performance in soccer can be challenging for players who do not contribute much directly to scoring goals, although innovation in collecting data on professional soccer matches could one day solve this problem. But as things stand today, it is difficult to quantify the effect a defensive midfielder like Kyle Beckerman has on the performance of Real Salt Lake, although the intuition of a soccer fan would say that his presence in the lineup has a tremendous impact on that team's success. As far as professional tennis is concerned though, I believe it shares more similarities with baseball than it does with American football and soccer. For one, the Association of Tennis Professionals (ATP), the world's governing body for men's professional tennis, compiles good data for the matches it sanctions, and this data is available to the public (Note: all stats and figures I cite have been pulled directly for the ATP's website, with the exception of information for which I specifically have cited another source).[8] In addition to having quality data, tennis is played in either singles or doubles. In this paper, my analysis will focus on ATP singles matches, so I will obviously not have an issue with conflating effects of different players within a team on that team's performance.

Because tennis does not have the large following that the major team sports do in America, I feel that I should give a brief overview of the structure of a tennis match. A point starts with serve. A serve which the opponent is unable to contact is called an ace. The server has two opportunities to make the serve into the diagonal service box. If he is unsuccessful in either attempt, he loses the point (missing both serves is called a double fault). Generally speaking, a point will end when the ball is hit into the net, lands outside of the court of play, or bounces twice. Although tennis has an odd scoring system (love, 15, 30, 40, deuce), in practice a game is won by the player who is first to reach four points, while winning by at least two points. The two players alternate serving between games. A set is won by the first player to win six games, but must also win by at least two games. A set can be won by any of the following scores: 6-0, 6-1, 6-2, 6-3, 6-4, 7-5, 7-6. The score of 7-6 is the one that stands out of the group because it seems to violate the requirement of winning by at least two games. If a set is tied with each player having won six games, it goes to a tiebreaker. In a tiebreaker, players alternate serving after odd-numbered points. It is over when one player reaches seven points, winning by at least two points. The match is won by the player winning two out of three sets, with a few exceptions. In the four major tournaments (Australian Open, French Open, Wimbledon, and US Open) and a few others, matches are best-of-five sets.

My dataset has been compiled from the ATP's website using a variety of metrics published by the organization. The ATP has listed these statistics for each of the top 200 players in the year-end rankings, dating back to 1991. As an avid follower of tennis and its history, I have reason to believe the general style of play has changed since 1991. For this reason, I have chosen divide and compare the two periods. While the year 2002 may seem arbitrary, it closely approximates the beginning of the era of Roger Federer and the end of the era of Pete Sampras. It is the beginning of what I view as the most modern era of professional tennis (Figures 1a and 1b).



(Figure 1a)



(Figure 1b)

Thus, my initial dataset will include year-end results from 2002 to 2014. When sifting through the ATP's data, I realized that I would not be able to use all of the two hundred observations for each year. Because a player's ranking in any given year is largely a function of his ranking the previous year, many players in the dataset remained in the top 200 in the world rankings who did not play enough matches in the year to do any sort of meaningful analysis. To give a hypothetical example, a player could have finished 2005 ranked 40th in the world. During the offseason, this hypothetical player could have sustained an injury which caused him to miss the bulk of the 2006 season. As a result, he may have only played in a half-dozen matches in 2006. Despite playing a limited number of matches, this player would be likely to finish in the top 200 of the 2006 year-end rankings. Additionally, it is also possible there could be an up-and-coming young player who reaches the top 200 in the world rankings while playing relatively few matches. Now it is obvious to me that these players need to be left out of the analysis. A player who participates in less than ten, or even less than twenty matches in a given year is much more likely to display match statistics that are the result of chance, rather than being truly indicative of the player's ability. By observing the data, I have judged it to be best to narrow my analysis to the top 50 players of each year's year-end rankings from 1991 to 2014. The overwhelming majority of the players who finished a year in the top 50 of the world rankings played at least 40 matches, with some playing as many as 100. My hope is that this will constitute enough matches to give nonrandom results, so I can draw meaningful inference from my analysis.

CHAPTER THREE

OVERVIEW OF THE ANALYSIS

I am looking to analyze my data by creating a model which relates a player's success to his skill level in various aspects of the game of tennis. I will divide the game into three skill categories: serving, returning, and court play. The last of these consists of both groundstrokes and net game. For determining a measure of success, my initial inclination was to use year-end rankings. However, this created a bit of a problem for me. First of all, the year-end ranking for the current year is largely a function of the year-end ranking of previous year. Furthermore, the formula used by the ATP for determining rankings is not made known to the public, and certain tournaments are weighted more heavily than others. My second thought was to use total wins for the year as my measure of success. But again, this approach has some obvious problems. Unlike team sports, there is no set schedule for professional tennis players. The Houston Astros are required to play 162 games during the season, unless of course they want to forfeit a game and have it recorded as a loss. They have no choice in the matter. On the other hand, professional tennis players have a choice as to which tournaments they enter. It is not uncommon to see star players sit out a tournament or two to rest as a part of their preparation for bigger tournaments. In addition, the nature of professional tennis tournaments lends itself to large discrepancies in total matches played amongst different players. As far as I am aware, all ATP-sanctioned tournaments are single-elimination. Many have either 64 or 128 entrants and have a bracket format similar to that of NCAA March Madness. Once a player loses a match, he is out of the tournament bracket. Because of this, one would expect for the very top players to play more matches over the course of the year than the less elite players, since the highest-ranking players are the individuals most likely to advance the furthest into tournaments.

For the reasons stated above, I ultimately decided to use winning percentage as my measure of player success and dependent variable for my regressions. Due to large differences in matches played, simply counting wins and losses will not make for good analysis. The same logic will apply to my measures of skill. Counting aces over the course of the year will not be a good measure of serving skill, but an aces per match statistic could possibly function as a good measure of serving skill. All stats used for my analysis will be rate stats, not counting stats. In Stata, I will be regressing player winning percentage on serving skill, returning skill, and court play skill. Now since there is no one metric to quantify these skills, I will run a number of regressions in an attempt to find the best proxies with the data that is available to me. Below I list the available stats by category.

Serving statistics:

- aces per match

- double faults per match
- percentage of first serves made
- percentage of first serve points won
- percentage of second serve points won
- percentage of serve games won
- percentage of break points saved

Returning statistics:

- percentage of first serve return points won
- percentage of second serve return points won
- percentage of break points won
- percentage of return games won

Other:

- winning percentage
- percentage of tiebreakers won
- tiebreakers played per match

Notably absent from this list is any potential measure of court play skill. Despite the fact the ATP compiles excellent data, it does not have available information regarding groundstrokes and net play. Because I believe it is necessary to control for court play skill in my model in order to avoid omitted variable bias, this lack of data presents a major problem for me. Therefore, I have created a statistic which I will call serve reliance. In my experience watching professional tennis, generally speaking, players with “big serves” hit many aces, don’t often break their opponent’s serve, play lots of tiebreakers, and most importantly, have weaker ground and net games, all else being equal. Although this is just anecdotal evidence, a few players immediately come to mind. This could be said about Andy Roddick, as he is most known for his big serve ability, and had weaker court play than other top 5 players when he was in his prime. The success of Roddick on any given day was heavily reliant on a good serving performance. But there are two other players whom I believe this concept is even more applicable. American John Isner and Croatian Ivo Karlovic are two extremely tall tennis players (6 ft. 10 in. and 6 ft. 11 in. respectively) who are known for enormous serve games and limited court mobility. Over the course of their careers, Isner has won just 11% of his return games, while Karlovic has won just 9% of his return games. To put this into perspective, Federer, Nadal, and Djokovic have won 27%, 33%, and 32% of their career return games.

I will calculate serve reliance by dividing tiebreakers played per match by the percentage of return games won. Players with a higher value for serve reliance will be considered to be more reliant on good serving performances relative to other players. When running my regressions, I expect the coefficient on this variable to be negative because I believe players who depend heavily on the performance of their serve will realize worse performances over the course of a given year than their peers who display similar characteristics in all other categories. The reasoning for my expectations is that I believe players will want to diversify their talent in a way much like investors want to diversify stocks. Players will want to be highly-skilled in many areas so they have a plan B or plan C if their go-to skill is not as effective for one reason or another in any one match. James Harden is a good example of this concept in professional basketball. If Harden is having an off-shooting night, his uncanny ability to draw fouls and get to the free throw line makes him able to compensate for his poor shooting. To bring it back to Isner and Karlovic, I get the sense that these two players do not have the same luxury. If their big serves happen to be neutralized by a good returner, their chances of winning that particular match are greatly diminished. There is one more thing regarding serve reliance that I need to note. Initially, I planned on calculating serve reliance by adding tiebreakers played per match and aces per match, then dividing the sum by percentage of return games won. I opted against this way of calculating the statistic because I did not want to overestimate the dependence on a quality serving performance for those players that hit a relatively high number of aces but also have good return games.

In order to simplify the analysis and to assure the reader I have accounted and controlled for all necessary factors, I am making the following assumptions:

- Aces and double faults are independent of the quality of the opponent. Since the serve is the first shot of the point, I believe that these stats will remain consistent for a given player across the entire spectrum of opponent talent. I realize in theory it is possible for this not to be the case. For example, if a player knows his opponent is an exceptionally good returner, he might take more risks on his serve in order to neutralize the strength of his opponent. This would seemingly lead to more aces and double faults. Given the data and nature of this paper, I am unable to compare aces and double faults for individual players while controlling for opponent quality. In order to do so, I would have to look at the match report for every single match for every single player for the past 13 years. This information is not readily available, thus I chose to make the above assumption.

- While I am assuming aces and double faults are independent of the opponent's skill level, the rest of the statistics are not likely to be independent of the quality of the opponent. All else being equal, playing against a good server will cause a player to win a smaller percentage of points and games as a returner. The reverse is also true. Playing a quality returner means the serve will be less effective than when playing against an average or mediocre returner. This makes it necessary for my model to control for the skill level of the player's opponents when predicting his year-end winning percentage. My assumption is that the quality of the opponents played will be reflected in the serving and returning skills statistics and the year-end winning percentage. For a given match, playing against a highly-skilled opponent will in general result in reduced skill stats, and in turn result in a greater probability of losing that match.

- Professional tennis tournaments are played primarily on one of three surfaces: hard courts, clay courts, and grass courts. Additionally, there are some matches played on indoor carpet courts contained in my dataset. The casual tennis fan will easily be able to recognize the fact that some players excel on a particular surface. The obvious example is Rafael Nadal when playing on clay courts. Nadal will go down in history as one of the greatest tennis players of all-time. He has achieved success on every surface, having won the Career Grand Slam (winning each of the four major tournaments). But of note is his record nine career victories at the French Open, the one major tournament played on clay. A similar point can be made about Roger Federer's success on grass courts. He too has won the Career Grand Slam, and a record 17 total major tournament titles. His success on grass courts stands out most, winning seven times at Wimbledon (the one major tournament played on grass), a record he shares with Pete Sampras. Just like in my above assumption regarding the quality of the opponent, I am assuming that a player's dominance on a particular surface will be reflected in the skill stats and winning percentage. To give an example, Nadal has won 43% of his career return games on clay courts, which is a staggeringly high figure. By comparison, he has won 27% of his return games on all other surfaces. Similar trends can be seen in other statistical categories for Nadal (see his player profile on www.atpworldtour.com). But Nadal's dominant skills on clay courts are reflected in his winning percentage. He was won 93% of his career matches on clay, compared to 77% of career matches played on all other surfaces. By this line of reasoning, I do not feel the need to explicitly control for the court surface in my model for predicting year-end winning percentage.

I am tempted to make the assumption that tiebreakers, since they are simply a microcosm of a set and match, will be won at higher rates by those players who win matches at higher rates. But I think it is also possible that tiebreaker performance is more related to serving skill. Therefore, I will resist the urge to make any assumptions about tiebreakers. Furthermore, because tiebreaker win-loss records are available in my dataset, I will be able to test empirically the relationship between match winning percentage and tiebreaker winning percentage.

CHAPTER FOUR

SUMMARY STATISTICS

Before running regressions in Stata, I wanted to view the summary statistics of my dataset to get an idea what kind of numbers I was working with. For clarity, here is a list of the variable names with their corresponding skill statistics:

- winpct: percentage of matches won
- tbrkmtch: tiebreakers played per match
- acemtch: aces per match
- reliance: serve reliance
- onesrvpct: percentage of first serves made
- onesrvwon: percentage of points won when making the first serve
- twosrvwon: percentage of points won when making the second serve
- oneretwon: percentage of points won when the opponent makes the first serve
- tworetwon: percentage of points won when the opponent makes the second serve
- retgamewon: percentage of return games won
- srvgamewon: percentage of serve games won
- tbrkpct: percentage of tiebreakers won

When observing the summary statistics (Table 1), I found it somewhat interesting that some of the skill statistics had rather large spreads, while others had relatively low spreads. Winning percentages in my dataset seem to be fairly spread out, with a high of 95% and a low around 37%. The sample mean was 61%. This is different from what we see in the team sports, where we expect the average winning percentage to be no more or no less than 50%. However, I am not at all surprised to see the average winning percentage for my sample to be above 50%. Because my analysis is restricted to the top 50 players in the year-end rankings, many of the matches played by players in the dataset were played against players not in the dataset. Additionally, since the players in the dataset are ranked higher than players not in the dataset, it should not be shocking that the higher ranked players win the majority of these matches, resulting in the 61% figure present in the summary statistics. The standard deviation for my sample was around 0.1, meaning about 95% of the player year-end winning percentages from the years 2002 to 2014 fell between 40% and 80%. Again, this result is not one I find to be surprising. Somewhat more interesting is the initial information I found regarding tiebreakers. Like winning percentages, the two tiebreaker metrics in my dataset were had fairly large spreads. For tiebreakers played per match, the range was from 0.17 to 1.1. Averaging more than a tiebreaker per match, when most matches are best two out of three sets, is incredibly high. I was also did not expect to find any observations with less than one tiebreaker played every five matches on average. Tiebreaker winning percentage and the serve reliance statistic I constructed also had relatively large variances. As I had anticipated from years of viewing professional tennis, there existed a large gap between the highest and lowest aces per match (20.6 vs. 0.6). The same could not be said of the remaining measures of serving skill, and the measures of returning skill, which all had less spread out observations.

I started a more complete analysis of the 1991-2001 data by observing the summary statistics from that period (Table 1). Upon an initial look at the numbers, the data appears to be awfully similar to the data from the 2002-2014 period. As a percentage of the sample mean from the more recent period, the sample means for ten of the twelve skill metrics in my dataset for 1991-2001 are nearly the same. The two exceptions are aces per match and the serve reliance metric I devised. From the former period to the latter, these two statistics saw their means increase by 17.7% and 22.8% respectively. This could potentially suggest a change in the relevance of these metrics over time, something which will be evaluated in my regression analysis of the former period. The most striking observation to be made when viewing this period’s summary statistics has to do with standard deviations of the metrics in the dataset. Nine of the twelve statistics in the sample have smaller standard deviations in the period that includes the 1990s than the period that includes the most recent years of professional tennis data. My first instinct tells me that this could be an indication of a widening gap between the very best players and the remaining athletes in the world's top 50, although other explanations could exist.

	2002-2014				1991-2001			
	mean	std. dev.	min	max	mean	std. dev.	min	max
winpct	0.601	0.101	0.367	0.953	0.613	0.087	0.407	0.923
tbrkmtch	0.447	0.137	0.170	1.100	0.414	0.115	0.050	1.120
acemtch	6.309	3.097	0.600	20.600	5.361	2.860	0.800	15.700
reliance	2.057	1.312	0.540	12.050	1.675	0.762	0.200	8.620
onesrvpct	0.609	0.044	0.500	0.740	0.590	0.049	0.460	0.780
onesrvwon	0.728	0.039	0.610	0.850	0.724	0.048	0.600	0.860
twosrvwon	0.522	0.025	0.450	0.600	0.507	0.022	0.440	0.570
oneretwon	0.305	0.031	0.210	0.450	0.304	0.026	0.220	0.380
tworetwon	0.505	0.030	0.400	0.610	0.519	0.025	0.430	0.600
retgamewon	0.242	0.050	0.090	0.390	0.263	0.044	0.130	0.390
srvgamewon	0.808	0.048	0.650	0.940	0.790	0.049	0.650	0.920
tbrkpct	0.537	0.113	0.000	0.857	0.539	0.106	0.150	1.000
	650 observations				550 observations			

(Table 1)

When I began researching this topic, I expected to have a model in which success (winning percentage) was a function of serving skill, returning skill, and court play skill. Some of my initial findings have caused me to reevaluate this model, and make an addition. Earlier, I noted that I was resisting the urge to make any assumptions about tiebreakers. Because they are simply a microcosm of the match as a whole, it seems intuitive to think that the best players who win a high percentage of their matches would be the same players who win a high percentage of their tiebreakers. As it turns out, this statement is not entirely true. While match and tiebreaker winning percentages do move together, the two metrics only have a moderately positive correlation (Figure 2).

Just like for the latter period, I wanted to look closer at the tiebreaker winning percentage before making any assumptions about their impact (or lack thereof) on overall player success (year-end winning percentage) in the 1991-2001 period. Like in my previous analysis, I find that in the chronologically earlier data, winning percentage and tiebreaker winning percentage move together (Figure 2). However, an even weaker correlation between the two exists than it did for the more recent data. Because of this I once again, despite tiebreakers simply being a microcosm of the tennis match as a whole, feel compelled to have a model which makes year-end winning percentage a function of serving skill, returning skill, court play skill, and tiebreaker skill.

	2002-2014		1991-2001	
	1	2	1	2
1. winpct	1.000	0.399	1.000	0.282
2. tbrkpct	0.399	1.000	0.282	1.000

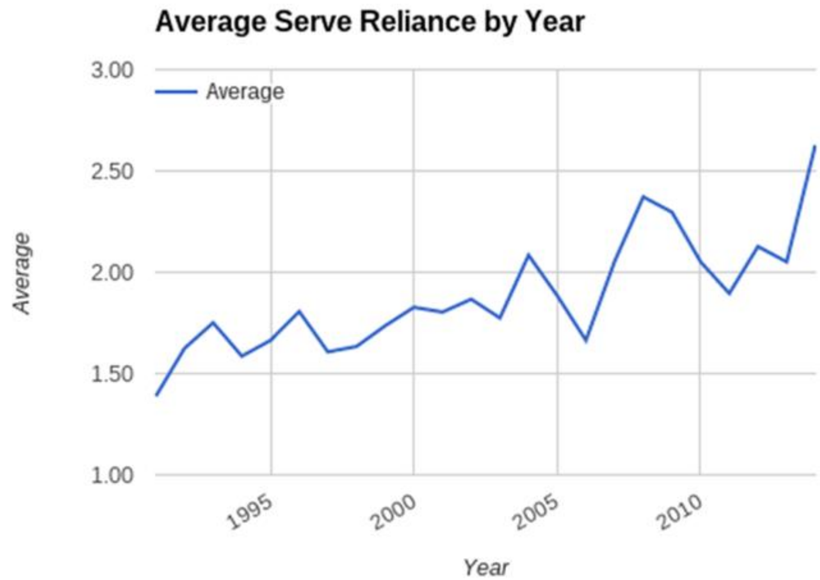
(Figure 2)

So while it may be true to say that in general top players perform well in tiebreakers, it would not be wise to say overall match success is a good predictor of success in tiebreakers. This leads me to believe that tiebreaker performance, in and of itself, is a skill. For this reason, I chose to alter my model to include tiebreaker skill, which is proxied by year-end tiebreaker winning percentage.

CHAPTER FIVE

RESULTS

Professional tennis looks much different today than it did fifty or a hundred years ago. I believe much of this can be attributed to significant advances in racket technology.9 Throughout much of the sport’s history, rackets were made out of laminated wood. The 1960s saw the introduction of steel and aluminum racket frames. By the early 1980s, graphite composites and other materials began to be used by racket producing companies. While composite racket frames are still the standard today, their quality has continued to improve greatly over the last three and a half decades. This gives me reason to believe it is possible that the skills with the greatest impact on winning percentage in the most recent years may not be the same skills that have always been most important for player success. When looking at the serve reliance statistic for the entire dataset (from 1991 to 2014), a noticeable trend emerges (Figure 3). Although a few outlier years exists, the average serve reliance for the top 50 players in each year’s rankings has drifted steadily upward, from less than one and a half in 1991 to more than two and a half in 2014.



(Figure 3)

I wanted to start my statistical analysis by looking at the obvious. It seems reasonable to say that players who win large percentages of their serve games, return games, and tiebreakers win large percentages of their matches. After all, doing so leads a player directly to victory. When using percentage of serve games won, percentage of return games won, and percentage of tiebreakers won, I find that all three have positive and statistically significant effects on year-end match winning percentage, and the model has a fairly large R-squared value (Table 2). As I would have expected, the coefficients for serve and return games won were much larger than the coefficient on tiebreakers (about ten times as great). This makes sense, at least in part, because players are guaranteed at least three serve and three return games per set, while the overwhelming majority of players average less than one tiebreaker per match. Although the magnitude of the effects of serve and return games were close, the coefficient on percentage of return games won is slightly higher. On average, a player who increases his percentage of serve games won by one percentage point will increase his winning percentage by 1.7 percentage points, whereas the expected effect of increasing percentage of return games won by one percentage point is a 1.5 percentage point increase in winning percentage. I feel the best way to look for significant differences between the two periods is to run the same regressions for the 1991-2001 period that I do for the most recent data. I ran the same initial regression I did for the 2002-2014 period (Table 2). The results from this regression are very similar to those of the other period. All three independent variables are still highly statistically significant and have positive coefficients of close to the same magnitude.

	2002-2014	1991-2001
winpct	coefficient	coefficient
srvgamewon	1.721	1.680
	(0.031)	(0.038)
retgamewon	1.509	1.420
	(0.029)	(0.042)
tbrkpct	0.198	0.175
	(0.012)	(0.015)
constant	-1.254	-1.182
	(0.028)	(0.037)
	R-squared = 0.883	R-squared = 0.816

Standard errors in parentheses.

(Table 2)

When adding serve reliance (Table 3) to this first model in order to control for the quality of a player's court game, I find that its coefficient is statistically and economically insignificant. As one could imagine, the addition of serve reliance, due to its insignificance, resulted in only the slightest of upticks in the adjusted R-squared value from the model for which it was omitted. Adding my serve reliance statistic to my model for the 1991-2001 period to control for the quality of a player's court game yields nearly identical results to when I first added it to my model for the 2002-2014 dataset (Table 3). I find that serve reliance is economically and statistically insignificant, in addition to have the opposite sign than what I would have expected.

	2002-2014	1991-2001
winpct	coefficient	coefficient
srvgamewon	1.708	1.658
	(0.034)	(0.062)
retgamewon	1.550	1.508
	(0.044)	(0.062)
tbrkpct	0.197	0.174
	(0.002)	(0.015)
reliance	0.002	0.007
	(0.002)	(0.004)
constant	-1.257	-1.199
	(0.028)	(0.038)
	R-squared = 0.883	R-squared = 0.817

Standard errors in parentheses.

(Table 3)

Although the variables used above explain a great deal of the variation in year-end winning percentage, they do not explain much in any practical sense. Of course winning more serve and return games helps a player win matches. That much is obvious. But it does not help a player to narrow down which specific skills are doing the most to help win matches. So in an attempt to address this dilemma, I chose to break up serving and returning skills into a few different categories. After breaking up the serving and return skills, I found that aces per match was a statistically insignificant metric for serving skill in both periods (Table 4). It was the case that every regression I ran that included aces per match gave me the result that the metric was insignificant, including regressions for which I use White standard errors to correct for any issues with heteroscedasticity. While I was not necessarily expecting this, I cannot say that I am altogether surprised at this finding. Because of limitations in my dataset, I am unable to net out aces from the percentage of first serve points won statistic. If I were able to do so, it is quite possible that aces would have a statistically significant effect on winning percentage. However, it is also possible that netting out aces would not lead to a statistically significant effect on winning percentage. From watching professional tennis, my gut tells me that aces and overall serving skill are correlated, but sheer volume of aces alone is probably not a good indicator of overall serving skill. Most aces come on first serves, so it is plausible that a player could have an above average first serve (and hit a high number of aces), yet have a mediocre second serve, which could result in just so-so overall serving skill. Because I continually found aces per match to be statistically insignificant, and I did not want to inflate the standard errors for any of my other estimators, I chose to omit aces per match from any of my final models.

	2002-2014	1991-2001
winpct	coefficient	coefficient
acemtch	0.003	-0.001
	(0.002)	(0.002)
reliance	-0.018	-0.005
	(0.004)	(0.007)
onesrvpct	0.550	0.420
	(0.053)	(0.050)
onesrvwon	1.385	1.642
	(0.137)	(0.112)
twosrvwon	1.167	1.429
	(0.109)	(0.101)
oneretwon	0.467	1.293
	(0.160)	(0.116)
tworetwon	1.156	1.224
	(0.125)	(0.107)
tbrkpct	0.199	0.123
	(0.020)	(0.018)
constant	-2.160	-2.628
	(0.124)	(0.092)
	R-squared = 0.727	R-squared = 0.747

Robust standard errors in parentheses.

(Table 4)

I then chose to use percentage of first and second serve points won as my measures of serving skill and percentage of first and second serve return points won as my measures of returning skill, controlling for court play and tiebreaker skills. The results show that all estimators are statistically significant and have the signs I would have expected (all are positive except for a negative sign on serve reliance). Lastly, I included percentage of first serves made in my model (Table 5). I was somewhat surprised to find this variable had a statistically and economically significant effect on year-end winning percentage. However, it does make sense that forcing your opponent to return more first serves will lead to better outcomes, provided the quality of the first serve is held constant. Therefore, I consider this final model to be my best model. I was interested to see the results of my final model for the former period indicate, that despite having the expected negative sign, the estimator for serve reliance was economically and statistically insignificant, again including when robust standard errors are used (Table 5). This result would appear to indicate that overreliance on good serving performances has a negative impact on a player's success in the most recent time period, while not having a consequential effect on performance in earlier years in the dataset.

	2002-2014	1991-2001
winpct	coefficient	coefficient
onesrvpct	0.557	0.417
	(0.056)	(0.050)
onesrvwon	1.558	1.600
	(0.079)	(0.070)
twosrvwon	1.126	1.435
	(0.106)	(0.098)
oneretwon	0.468	1.299
	(0.100)	(0.113)
tworetwon	1.112	1.229
	(0.111)	(0.106)
tbrkpct	0.198	0.123
	(0.019)	(0.018)
reliance	-0.016	-0.006
	(0.003)	(0.007)
constant	-2.230	-2.608
	(0.084)	(0.091)
	R-squared = 0.726	R-squared = 0.746

Robust standard errors in parentheses.

(Table 5)

Because of its relevance to the real world implications of my analysis, I performed an F-test to see if my coefficients are statistically different from one another. I find that they are for the 2002-2014 period. If my model is to be trusted, I can conclude that first serve skill has a greater impact on winning than second serve skill, and second return skill has a greater impact on winning than first return skill for this period (Figure 4). Doing the same F-test for the 1991-2001 period I believe led to the most interesting result that reveals a significant distinction between the two periods. The coefficients on percentage of first serve points won and percentage of second serve points are not statistically different from one another (Figure 4). This means that an improvement in first serve skill does not have a larger impact on winning percentage than a proportional improvement in second serve skill, which was found in the data for the latter period. The same can be said for the coefficients on percentage of first serve return points won and percentage of second serve return points won. While in the more recent years, second serve return skill had a larger impact on a player's success, the impact of the first and second serve return skill are statistically the same in the earlier years. This could possibly be an indication that there may exist a single, better and more inclusive metric for each of the two skills (serve and return) in this period.

2002-2014	1991-2001
onesrvwon=twosrvwon	
Prob > F = 0.0067	Prob > F = 0.1551
oneretwon=tworetwon	
Prob > F = 0.0002	Prob > F = 0.6089

(Figure 4)

CHAPTER SIX

GOING FORWARD

Given the limited amount of research done until now in analytics for professional tennis, it is my belief that there exists enormous potential for more work in this field. Although numbers gurus will always be valuable for the purpose of moving knowledge forward, Daryl Morey notes that “the age of the irreplaceable analyst no longer exists, if it ever did.”[10] According to the general manager of the Houston Rockets, the way to better analysis is better data.

> “Raw numbers, not the people and programs that attempt to make sense of them. Many organizations have spent the last few years hiring top analysts based on the belief that they create differentiation. Smart companies such as Google believe they need savants to crunch those numbers and find the connections that regular humans could not. But my experience, and what I’m hearing from more organizations (sports and non), shows that real advantage comes from unique data that no one else has.”

Morey goes on the give a relevant example as it relates to the NBA:

> “Many teams in the NBA track data for their own team such as how often a player on defense challenges shots. When tracked for your own team, this information can be useful to add accountability to the important things a

> coach is trying to emphasize to win games and to improve players on the margin by increasing their effort on challenging shots. The data does not offer significant competitive leverage, however, until you track the data *for the entire league*. Only with the league-wide data can you tell if your players are creating an advantage relative to others in the league on shot challenges (higher leverage) or even more important, identify players you may want to acquire who challenge shots extremely well (highest leverage). Without the context of the entire league, it is very hard to use data in any meaningfully competitive way. Tracking data for the whole league across multiple dimensions is a significant task but very worth it. For obvious reasons, I cannot reveal what data the Houston Rockets track but to track the significant data we gather we use a very large set of temporary labor that helps us develop these data sets that we hope will create an advantage over time. To be sure, you need strong analysts (and we have many) to then work with this data, but the leverage comes not from the analysis but from having the data that others do not."

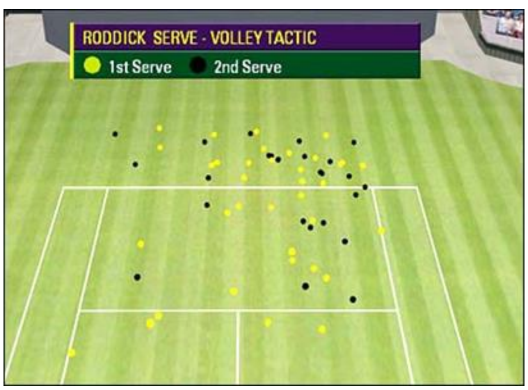
As I briefly mentioned earlier, it did not take long for the rest of Major League Baseball to catch up to the Oakland A's. Once the high revenue teams, such as the Boston Red Sox and New York Yankees, found out what was going on in Oakland, they too hired smart analysts, and Oakland's advantage quickly eroded. They only way for Oakland to achieve sustained success with its limited payroll would have been to obtain better data than any of its competitors.

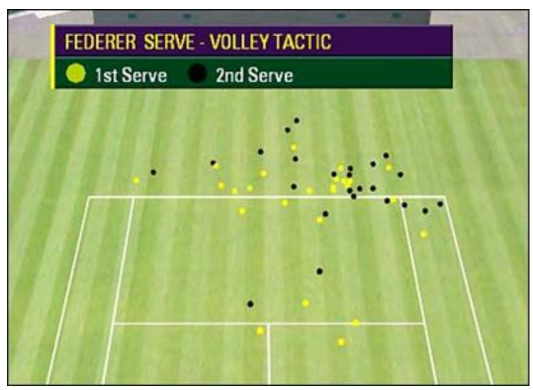
The most likely place from which better data will come in professional tennis is via Hawk-Eye technology.[11] Developed by Dr. Paul Hawkins in the United Kingdom, the technology was first used for cricket in the early 2000s, but is now used in a variety of professional sports (including tennis, soccer, and Australian rules football). The technology works by employing the principles of triangulation and a number of high-speed cameras (ten in professional tennis) to track the path of the ball. The graphics produced by the system allow for information to be provided to judges and officials, television viewers, and coaches almost instantaneously. Although the technology is not perfect, it is able to judge the location of the ball with a fair amount of precision, having a margin of error of just 5 millimeters.

The Hawk-Eye technology has been used for quite some time now to analyze individual matches, particularly important matches with large television viewership in the major tournaments. For example, BBC Sport has an in-depth analysis of the 2005 Wimbledon men's singles final in which Roger Federer defeated Andy Roddick 6-2, 7-6 (7-2), 6-4.[12] The data indicates that both Federer and Roddick attempted a fairly even mix of first serves to their opponent's backhand and forehand sides (Table 6). However, on second serve attempts, both players primarily attack their opponent's backhand side (which would be considered the norm in professional tennis). Data from player third shots within a rally (Figures 5a and 5b) seems to suggest that Roddick was somewhat more aggressive in his attempt to get to the net to hit volleys (42 approach shots), while Federer was more content to play from the baseline (25 approach shots).

2005 Wimbledon Men's Final: Serves to Opponent's Backhand					
Roddick	**Deuce court**	**Ad court**			
1st	36%	48%			
2nd	71%	88%			
Federer	**Deuce court**	**Ad court**			
1st	57%	48%			
2nd	85%	80%			

(Table 6)





(Figure 5a)

(Figure 5b)

At this juncture, the Hawk-Eye technology is not able to be used for comprehensive analysis. In order to use it in a manner similar to which I did for the rest of my paper, the information would need to be available for all matches involving all of the top professional players. While some interesting analysis of individual matches has been done, complete datasets are necessary to make meaningful inferences about winning percentage. If and when these datasets are compiled, my instincts as a player and fan tell me that the information regarding serve and groundstroke court depth will be most relevant. For a given point, playing the ball deeper in the service box for a serve and deeper in the court for a groundstroke means that the opponent is more likely to be in a defensive position and therefore less likely to win the point. It seems reasonable to assume that players with deeper average serve and shot depth are more frequently putting their opponents in defensive positions and more likely to be successful over the course of a year's worth of matches. Although less certain, I also think the arc of a player's groundstrokes could potentially play a role in year-end winning percentage. Shots hit with larger amounts of topspin tend to travel in more of an arched trajectory while shots with less topspin travel in a relatively flat trajectory. The Hawk-Eye technology is capable to giving this information. My experience tells me that shots with heavy topspin (like Nadal's forehand) are more difficult to deal with as a player. It is possible that greater average shot arc leads to greater winning percentages.

Academic research on mixed strategy equilibrium also gives some idea of which direction future tennis analysis could go given complete Hawk-Eye datasets. Chiappori, Levitt, and Groseclose tested mixed strategy equilibria in game theory by studying penalty kicks in professional soccer.[13] Their data from the major leagues in Europe made for good analysis because incentives are properly aligned. The top leagues in Europe are the best in the world. Millions of dollars are at stake, giving players and coaches ample reason to invest the time necessary to fully understand the game, its strategies, and to choose optimally. In the game of penalty kicks, for all intents and purposes, the players move simultaneously. Kickers have the choice of kicking their natural side or opposite side, while goalkeepers have the choice to dive to their natural side or opposite. Given payoffs conditioned upon the strategies of other players, the authors conclude that the professional optimize in a way consistent with what the theory suggests. Walker and Wooders do similar analysis on serve data from Wimbledon.[14] While they find that players do mix between serving to their opponent's forehand and backhand side, doing a runs test reveals professional tennis players choose strategies inconsistent with the theory. There exists serial dependence in player serve strategies. As opposed to mixing strategies randomly, players too often alternate between serves to the opponent's forehand and the opponent's backhand. Given the fact that this information has been available now since the early 2000s, is it possible that players have corrected any inefficiencies in their serve strategies, and furthermore, do strategize optimally when it comes to hitting groundstrokes?

I believe it could be possible that playing an unpredictable mix of heavily arched and flat shots will aid in success. Additionally, general tactics and strategic execution are vital to a player's success. Frank Giampaolo places an emphasis on court positioning and shot selection. “Being in the right place at the right time to maximize success” and “executing patterns and plays at the appropriate time” are vital.[7] The Hawk-eye technology has the ability to allow for in-depth exploration of optimal court positioning and shot selection strategies. Having more available data through Hawk-Eye will help to answer these questions and give players, coaches, and fans a better overall understanding of what elements of the game of professional tennis have the greatest impact on player success.

REFERENCES

1. James, Bill. *Historical Baseball Abstract*. New York: Villard, 1988. Print.
2. Lewis, Michael. *Moneyball: The art of winning an unfair game*. WW Norton & Company, 2004.
3. Hakes, Jahn K., and Raymond D. Sauer. "An Economic Evaluation of the Moneyball Hypothesis." *Journal of Economic Perspectives* 20.3: 181.
4. Oliver, Dean. Basketball on Paper. 1st ed. Dulles: Brassey's, 2004. Print.
5. Pagels, Jim. "Why Is Tennis So Far Behind Other Sports In Data Analytics?" *Sportsmoney*. Forbes, 3 Mar. 2015. Web.
6. "2001 Final Standings." MLB.com. Major League Baseball. Web.
7. Giampaolo, Frank. Championship Tennis. Champaign: Human Kinetics, 2013. Print.
8. "ATP MatchFacts." *ATP World Tour*. Association of Tennis Professionals. Web. <http://www.atpworldtour.com/Rankings/Top-Matchfacts.aspx>.
9. "The Quest for the Tennis Sweet Spot." BBC Sport. BBC, 14 Sept. 2005. Web.

10. Morey, Daryl. "Better Data, Not Better Analysis." Harvard Business Review. Harvard Business Publishing, 8 Aug. 2011. Web.
11. "Electronic Line Calling." Hawk-Eye. Hawk-Eye Innovations Ltd. Web.
12. "Hawk-Eye Analysis: Federer v Roddick." BBC Sport. BBC, 4 July 2005. Web.
13. Chiappori, P.-A., S. Levitt, and T. Groseclose. "Testing Mixed-Strategy Equilibria When Players Are Heterogeneous: The Case of Penalty Kicks in Soccer." American Economic Association (2002): 1138-151. Print.
14. Walker, Mark, and John Wooders. "Minimax Play at Wimbledon." The American Economic Review (2001): 1521-538. Print.